Wikipedia Data Crunch

Using Big Data Tools to
Analyze Our Collective Curiosity

Analytics Data Used

- **Clickstream Data**
  https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream
  - Data Explanation: https://dumps.wikimedia.org/other/clickstream/readme.html

- **Pageview Data**
  https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Pageviews
  - Data Explanation: https://dumps.wikimedia.org/other/pageviews/readme.html

- **Revision History**
  https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Edits/Mediawiki_history_dumps#Technical_Documentation
  - Data Explanation: https://dumps.wikimedia.org/other/mediawiki_history/readme.html

Question 1:

Which English wikipedia article got the most traffic on October 20, 2020?

Methodology

1. Gather pageview data for the given time span.
2. Aggregate pageviews for each page.
3. Determine non-trivial highest traffic page.

Assumptions

- Use the time span for which the three anglophone countries may be using Wikipedia on 10/20/2020.
  - This span is discussed in a later question.

- Wikipedia's landing page (Main Page), Special: Search, and Hyphen-Minus are considered trivial in pageview ranking.
  - For a discussion on why hyphen-minus appears at the top of the pageview rankings, check here: https://www.reddit.com/r/wikipedia/comments/fnwfop/aside_from_the_main_page_the_most_viewed_page_on/

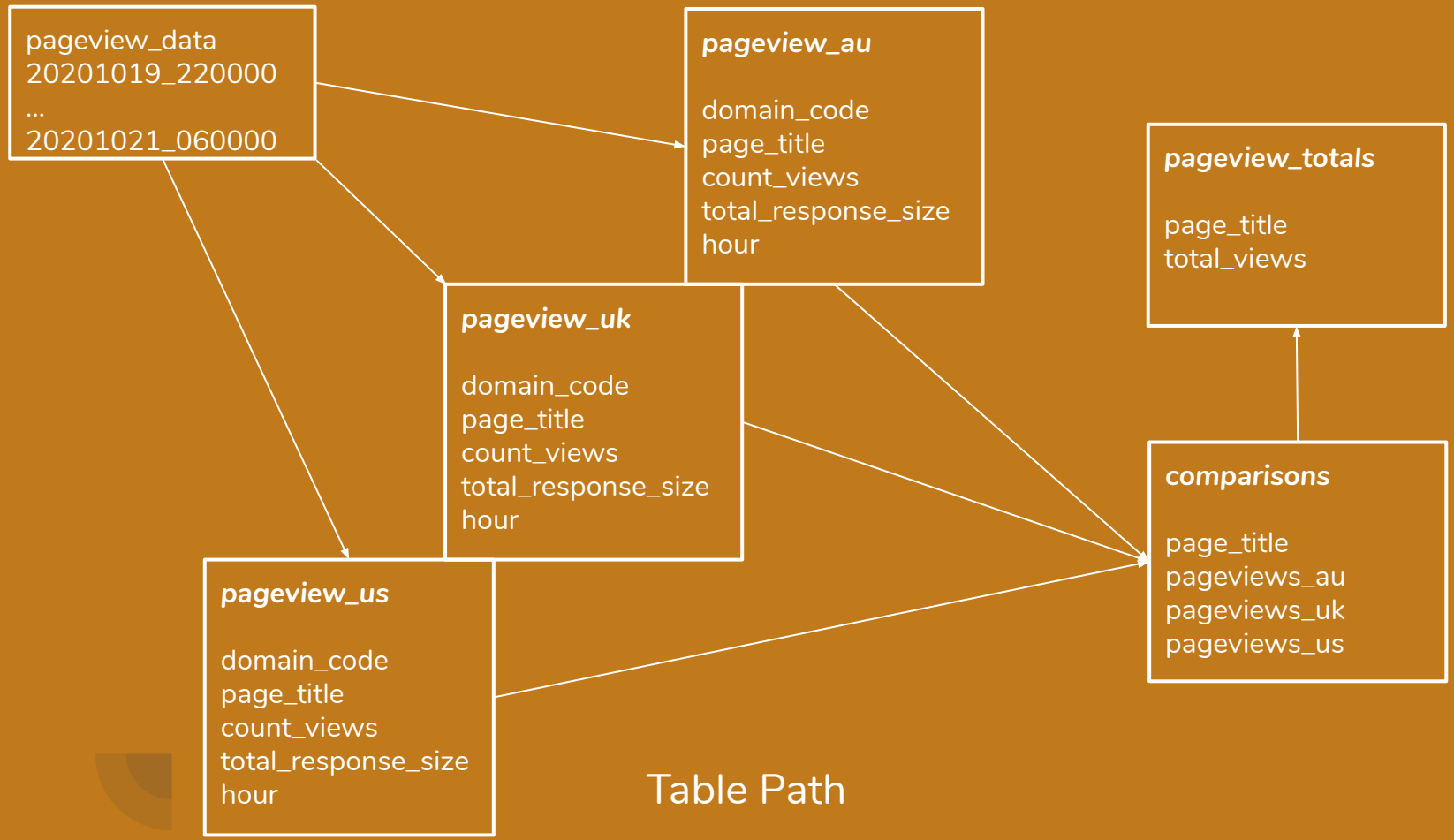

pageview_data
20201019_220000
...
20201021_060000
pageview_au
domain_code
page_title
count_views
total_response_size
hour
pageview_totals
page_title
total_views
pageview_uk
domain_code
page_title
count_views
total_response_size
hour
comparisons
page_title
pageviews_au
pageviews_uk
pageviews_us
pageview_us
domain_code
page_title
count_views
total_response_size
hour

Table Path

```
0: jdbc:hive2://> SELECT *
. . . . . . . . > FROM total_views
. . . . . . . . > ORDER BY total_views DESC
. . . . . . . . > LIMIT 20;
```

SQL Query

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.7 sec   HDFS Read: 157139486 HDFS Write: 769 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 700 msec
OK
+-----------------------------------------+--------------------------+
|          total_views.page_title         | total_views.total_views  |
+-----------------------------------------+--------------------------+
| Main_Page                               | 7986969                  |
| Special:Search                          | 2057872                  |
| -                                       | 729654                   |
| Jeffrey_Toobin                          | 418944                   |
| Sokushinbutsu                           | 300123                   |
| The_Haunting_of_Bly_Manor               | 267252                   |
| C._Rajagopalachari                      | 218431                   |
| Chicago_Seven                           | 218337                   |
| Robert_Redford                          | 209727                   |
| Bible                                   | 206792                   |
| Jeff_Bridges                            | 179674                   |
| The_Trial_of_the_Chicago_7              | 165882                   |
| Abbie_Hoffman                           | 162057                   |
| Deaths_in_2020                          | 154081                   |
| Harshad_Mehta                           | 145773                   |
| QAnon                                   | 137275                   |
| Hunter_Biden                            | 132765                   |
| Joe_Biden                               | 131127                   |
| 2016_United_States_presidential_election | 130401                   |
| Tom_Hayden                              | 128877                   |
+-----------------------------------------+--------------------------+
20 rows selected (16.611 seconds)
```

Query results

Question 2:

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

Methodology

1. Determine the number of linked views for each page (internal and external).

2. Determine the number of internal links that a page spawns.

3. Divide number of internal links by total linked views.

Assumptions

- Some pages receive fewer linked views than the number of link clicks they generate. This could be because one viewer is clicking on multiple links in a page, or it could be incompleteness of the data.
  - For analysis purposes anything with a ratio higher than 300% (i.e. three link clicks for every viewer) will be discarded.
  - A better analysis methodology would be to gather historic pageview data for the time period covered in the clickstream data, but here we'll work with the data given.

- Since very low numbers of total traffic will cause artificially high ratios, only pages with more than 3000 total traffic links will be considered.

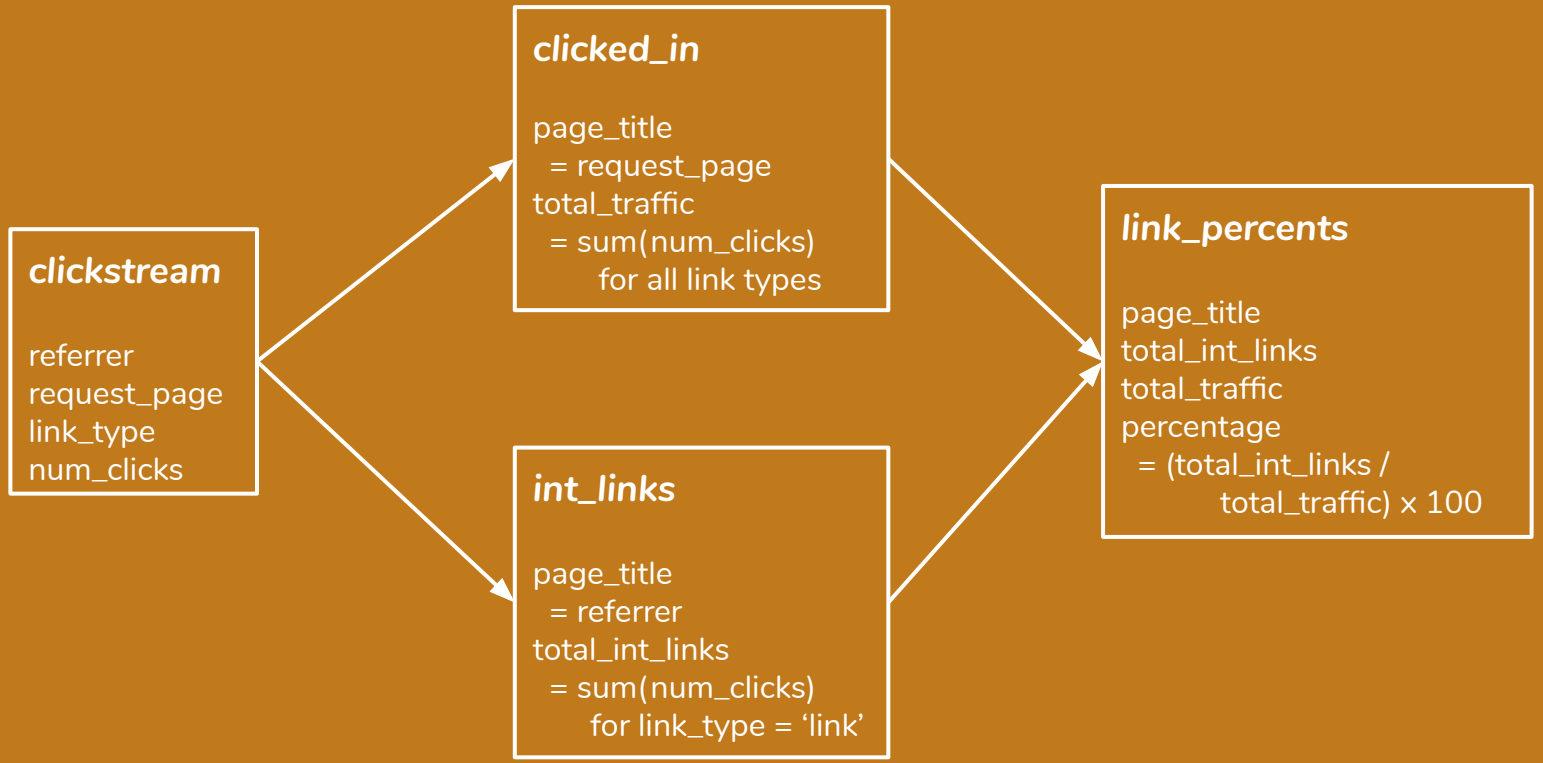

clickstream
referrer
request_page
link_type
num_clicks
clicked_in
page_title
= request_page
total_traffic
= sum(num_clicks)
for all link types
int_links
page_title
= referrer
total_int_links
= sum(num_clicks)
for link_type = 'link'
link_percents
page_title
total_int_links
total_traffic
percentage
= (total_int_links /
total_traffic) x 100

Table Path

```
0: jdbc:hive2://> SELECT *
. . . . . . . . > FROM link_percents
. . . . . . . . > WHERE percentage <= 300
. . . . . . . . >   AND total_traffic > 3000
. . . . . . . . > ORDER BY percentage DESC
. . . . . . . . > LIMIT 20;
```

SQL Query

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.86 sec   HDFS Read: 68589032 HDFS Write: 1525 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 860 msec
OK
```

link_percents.page_title	link_percents.int_links	link_percents.total_traffic	link_percents.percentage
List_of_World_War_II_battles	40435	13767	293.71
List_of_wars_involving_the_Ottoman_Empire	9168	3124	293.47
List_of_people_who_died_by_hanging	32564	11316	287.77
List_of_interments_at_Forest_Lawn_Memorial_Park_(Hollywood_Hills)	43431	15098	287.66
List_of_Subaru_vehicles	9056	3160	286.58
List_of_land_vehicles_of_the_U.S._Armed_Forces	9287	3316	280.07
List_of_current_NFC_team_rosters	27445	9808	279.82
List_of_premature_professional_wrestling_deaths	113946	40851	278.93
List_of_armoured_fighting_vehicles_by_country	13639	4930	276.65
List_of_wars:_before_1000	15688	5679	276.25
List_of_German_military_equipment_of_World_War_II	45359	16476	275.30
List_of_critically_endangered_mammals	18398	6706	274.35
List_of_rotorcraft	9527	3494	272.67
List_of_second-generation_National_Basketball_Association_players	14948	5511	271.24
List_of_anti-aircraft_weapons	14907	5508	270.64
British_Commonwealth_armoured_fighting_vehicles_of_World_War_II	9705	3588	270.48
List_of_child_abuse_cases_featuring_long-term_detention	34603	12798	270.38
List_of_baseball_players_who_died_during_their_careers	24378	9027	270.06
List_of_pornographic_performers_by_decade	467454	173587	269.29
Types_of_swords	51361	19095	268.98

```
20 rows selected (13.534 seconds)
```

Query results

Question 3:

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

Methodology

1. Gather all pages linked to from the “home page”.
2. Join the click_stream table to the link_percentages table.
3. Order the linked pages by their percentage of internal links.
4. Make the top page our new “home page”.
5. Repeat.

Assumptions

- Again, low total traffic will result in artificially high link to viewer ratios, but by ordering the results by their link percentage and limiting the query to the first 20 results we should avoid any trivially high-ratio pages.

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 80.75 sec   HDFS Read: 1491015395 HDFS Write: 3877 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.2 sec   HDFS Read: 12794 HDFS Write: 873 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 22 seconds 950 msec
OK
+-------------------------------------+----------------+
|              request                | lp.percentage  |
+-------------------------------------+----------------+
| Eagles_(band)                       | 96.62          |
| Jethro_Tull_(band)                  | 87.02          |
| The_Twilight_Zone_(1959_TV_series)  | 79.53          |
| Steely_Dan                          | 79.13          |
| American_Horror_Story:_Hotel        | 74.29          |
| Cameron_Crowe                       | 67.32          |
| Desperado                           | 64.56          |
| Anton_LaVey                         | 63.33          |
| Hotel_California_(disambiguation)   | 61.80          |
| John_Fowles                         | 61.64          |
| Mercedes-Benz                       | 57.12          |
| The_Big_Lebowski                    | 55.92          |
| 20th_Annual_Grammy_Awards           | 54.60          |
| Close_Encounters_of_the_Third_Kind  | 53.43          |
| The_Royal_Scam                      | 52.34          |
| Grammy_Award_for_Record_of_the_Year | 49.86          |
| Taxi_Driver                         | 49.71          |
| Church_of_Satan                     | 44.97          |
| Hotel_California_(Eagles_album)     | 44.90          |
| Private_Times...and_the_Whole_9!    | 44.33          |
+-------------------------------------+----------------+
20 rows selected (45.603 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . > FROM click_stream
. . . . . > JOIN link_percents AS lp
. . . . . >   ON request = lp.page_title
. . . . . > WHERE prev = 'Hotel_California'
. . . . . > ORDER BY lp.percentage DESC
. . . . . > LIMIT 20;
```

Step 1: Hotel_California

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 79.81 sec   HDFS Read: 1491015356 HDFS Write: 8190 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.2 sec   HDFS Read: 17107 HDFS Write: 773 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 22 seconds 10 msec
OK
+-----------------------------------+---------------+
|              request              | lp.percentage |
+-----------------------------------+---------------+
| Eagles_discography                | 125.90        |
| The_Rolling_Stones                | 122.16        |
| Journey_(band)                    | 117.21        |
| The_Beach_Boys                    | 116.20        |
| Aerosmith                         | 112.27        |
| Deep_Purple                       | 111.90        |
| The_Doors                         | 106.48        |
| Pink_Floyd                        | 106.31        |
| Yes_(band)                        | 106.28        |
| Earth,_Wind_&_Fire                | 105.49        |
| Led_Zeppelin                      | 104.30        |
| Emerson,_Lake_&_Palmer            | 103.93        |
| The_Beatles                       | 101.66        |
| Fleetwood_Mac                     | 98.28         |
| Guns_N'_Roses                     | 97.93         |
| Grammy_Award_for_Album_of_the_Year | 97.60         |
| Queen_(band)                      | 95.19         |
| Bee_Gees                          | 89.11         |
| Crosby,_Stills,_Nash_&_Young      | 88.70         |
| The_Byrds                         | 83.16         |
+-----------------------------------+---------------+
20 rows selected (46.121 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . > FROM click_stream
. . . . . > JOIN link_percents AS lp
. . . . . >   ON request = lp.page_title
. . . . . > WHERE prev = 'Eagles_(band)'
. . . . . > ORDER BY lp.percentage DESC
. . . . . > LIMIT 20;
```

Step 2: Eagles_(band)

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 80.12 sec   HDFS Read: 1491015421 HDFS Write: 3066 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.26 sec   HDFS Read: 11983 HDFS Write: 970 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 22 seconds 380 msec
OK
+---------------------------------------+----------------+
|                request                | lp.percentage  |
+---------------------------------------+----------------+
| Olivia_Newton-John_albums_discography | 154.61         |
| Electric_Light_Orchestra_discography  | 146.68         |
| Earth,_Wind_&_Fire_discography        | 144.03         |
| Linda_Ronstadt_discography            | 133.11         |
| Led_Zeppelin_discography              | 127.62         |
| Garth_Brooks_discography              | 107.38         |
| Eagles_(band)                         | 96.62          |
| Don_Henley_discography                | 86.19          |
| Eagles_(box_set)                      | 66.37          |
| Selected_Works:_1972–1999             | 55.58          |
| The_Very_Best_of_the_Eagles           | 54.93          |
| Eagles_Live                           | 54.11          |
| Eagles_Greatest_Hits,_Vol._2          | 53.78          |
| On_the_Border                         | 51.98          |
| Adult_Contemporary_(chart)            | 50.01          |
| The_Very_Best_Of_(Eagles_album)       | 46.34          |
| One_of_These_Nights                   | 46.29          |
| Hotel_California_(Eagles_album)       | 44.90          |
| Farewell_1_Tour:_Live_from_Melbourne  | 44.84          |
| Hell_Freezes_Over                     | 43.70          |
+---------------------------------------+----------------+
20 rows selected (47.444 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'Eagles_discography'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 3: Eagles_discography

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 77.98 sec   HDFS Read: 1491015668 HDFS Write: 2819 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.26 sec   HDFS Read: 11736 HDFS Write: 1015 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 20 seconds 240 msec
OK
+-------------------------------------------------+----------------+
|                     request                     | lp.percentage  |
+-------------------------------------------------+----------------+
| Olivia_Newton-John_singles_discography          | 78.57          |
| One_Woman's_Live_Journey                        | 76.09          |
| Xanadu_(soundtrack)                             | 57.34          |
| Highlights_from_The_Main_Event                  | 56.36          |
| Clearly_Love                                    | 49.88          |
| Love_Performance                                | 48.62          |
| Don't_Stop_Believin'_(album)                    | 46.97          |
| Olivia's_Greatest_Hits_Vol._2                   | 45.79          |
| Grease:_The_Original_Soundtrack_from_the_Motion_Picture | 45.44          |
| Back_with_a_Heart                               | 44.20          |
| Long_Live_Love_(album)                          | 41.95          |
| Back_to_Basics:_The_Essential_Collection_1971-1992 | 41.43          |
| Olivia_Newton-John                              | 40.47          |
| Two_of_a_Kind_(soundtrack)                      |
| Gaia:_One_Woman's_Journey                       |
| Making_a_Good_Thing_Better                      |
| Physical_(album)                                |
| Come_On_Over_(Olivia_Newton-John_album)         |
| Have_You_Never_Been_Mellow                      |
| If_You_Love_Me,_Let_Me_Know                     |
+-------------------------------------------------
20 rows selected (45.629 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'Olivia_Newton-John_albums_discography'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 4: Olivia_Newton-John_albums_discography

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 79.99 sec   HDFS Read: 1491015681 HDFS Write: 4016 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.5 sec   HDFS Read: 12933 HDFS Write: 991 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 22 seconds 490 msec
OK
+-------------------------------------------------+----------------+
|                     request                     | lp.percentage  |
+-------------------------------------------------+----------------+
| Olivia_Newton-John_albums_discography           | 154.61         |
| Xanadu_(soundtrack)                             | 57.34          |
| Clearly_Love                                    | 49.88          |
| Don't_Stop_Believin'_(album)                    | 46.97          |
| Olivia's_Greatest_Hits_Vol._2                   | 45.79          |
| Grease:_The_Original_Soundtrack_from_the_Motion_Picture | 45.44          |
| Tied_Up                                         | 45.38          |
| Back_with_a_Heart                               | 44.20          |
| After_Dark_(Andy_Gibb_album)                    | 42.12          |
| Long_Live_Love_(album)                          | 41.95          |
| Back_to_Basics:_The_Essential_Collection_1971-1992 | 41.43          |
| Olivia_Newton-John                              | 40.47          |
| Two_of_a_Kind_(soundtrack)                      | 40.45          |
| Gaia:_One_Woman's_Journey
| Making_a_Good_Thing_Better
| Physical_(album)
| Come_On_Over_(Olivia_Newton-John_album)
| Have_You_Never_Been_Mellow
| If_You_Love_Me,_Let_Me_Know
| Totally_Hot
+-----------------------------------------
20 rows selected (46.544 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'Olivia_Newton-John_singles_discography'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 5: Olivia_Newton-John_singles_discography

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 78.92 sec   HDFS Read: 1491015434 HDFS Write: 1601 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.18 sec   HDFS Read: 10518 HDFS Write: 987 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 21 seconds 100 msec
OK
+-------------------------------------------------+----------------+
|                     request                     | lp.percentage  |
+-------------------------------------------------+----------------+
| A_Box_of_Their_Best                             | 88.92          |
| Four_Light_Years                                | 88.22          |
| Electric_Light_Orchestra                        | 87.41          |
| The_Tubes                                       | 64.49          |
| Discovery_(Electric_Light_Orchestra_album)      | 51.67          |
| Time_(ELO_album)                                | 45.26          |
| Xanadu_(film)                                   | 45.22          |
| Olivia_Newton-John                              | 40.47          |
| Physical_(album)                                | 38.93          |
| Totally_Hot                                     | 37.02          |
| Xanadu_(Olivia_Newton-John_and_Electric_Light_Orchestra_song) | 35.11          |
| All_Over_the_World_(Electric_Light_Orchestra_song) | 32.39          |
| Suddenly_(Olivia_Newton-John_and_Cliff_Richard_song) | 32.08          |
| Musical_film                                    | 30.82          |
| Cliff_Richard                                   | 30.30          |
| Magic_(Olivia_Newton-John_song)                 | 30.29          |
| I'm_Alive_(Electric_Light_Orchestra_song)       | 29.18          |
| James_Newton_Howard                             | 28.56          |
| Gene_Kelly                                      | 28.17          |
| Don't_Walk_Away_(Electric_Light_Orchestra_song) | 26.43          |
+-------------------------------------------------+----------------+
20 rows selected (46.459 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . > FROM click_stream
. . . . . > JOIN link_percents AS lp
. . . . . >   ON request = lp.page_title
. . . . . > WHERE prev = 'Xanadu_(soundtrack)'
. . . . . > ORDER BY lp.percentage DESC
. . . . . > LIMIT 20;
```

Step 6: Xanadu_(soundtrack)

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 77.32 sec   HDFS Read: 1491015434 HDFS Write: 786 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.34 sec   HDFS Read: 9703 HDFS Write: 282 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 19 seconds 660 msec
OK
+--------------------------+----------------+
|          request         | lp.percentage  |
+--------------------------+----------------+
| Four_Light_Years         | 88.22          |
| Electric_Light_Orchestra | 87.41          |
| Xanadu_(soundtrack)      | 57.34          |
| ELO's_Greatest_Hits      | 35.19          |
| Doin'_That_Crazy_Thing   | 15.46          |
+--------------------------+----------------+
5 rows selected (46.131 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'A_Box_of_Their_Best'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 7: A_Box_of_Their_Best

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 77.8 sec   HDFS Read: 1491015383 HDFS Write: 701 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.2 sec   HDFS Read: 9588 HDFS Write: 203 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 20 seconds 0 msec
OK
+--------------------------+----------------+
|          request         | lp.percentage  |
+--------------------------+----------------+
| Electric_Light_Orchestra | 87.41          |
| Xanadu_(soundtrack)      | 57.34          |
| Time_(ELO_album)         | 45.26          |
+--------------------------+----------------+
3 rows selected (46.852 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'Four_Light_Years'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 8: Four_Light_Years

```
Stage-Stage-1: Map: 7  Reduce: 6   Cumulative CPU: 78.5 sec   HDFS Read: 1491015499 HDFS Write: 6230 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.39 sec   HDFS Read: 15147 HDFS Write: 772 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 20 seconds 890 msec
OK
+----------------------------------------+---------------+
|                request                 | lp.percentage |
+----------------------------------------+---------------+
| Electric_Light_Orchestra_discography   | 146.68        |
| Elo                                    | 144.61        |
| The_Flaming_Lips                       | 108.35        |
| Led_Zeppelin                           | 104.30        |
| Emerson,_Lake_&_Palmer                 | 103.93        |
| The_Beatles                            | 101.66        |
| Fleetwood_Mac                          | 98.28         |
| The_Moody_Blues                        | 97.88         |
| The_Pussycat_Dolls                     | 96.94         |
| Queen_(band)                           | 95.19         |
| Take_That                              | 89.16         |
| Supertramp                             | 88.70         |
| Super_Furry_Animals                    | 78.02         |
| ELO_Part_II                            | 77.29         |
| Daft_Punk                              | 72.24         |
| The_Move                               | 71.27         |
| Bob_Marley                             | 70.90         |
| List_of_Electric_Light_Orchestra_members | 69.23       |
| Wizzard                                | 63.70         |
| Traveling_Wilburys                     | 60.86         |
+----------------------------------------+---------------+
20 rows selected (47.42 seconds)
```

```
0: jdbc:hive2://> SELECT request, lp.percentage
. . . . . . . . > FROM click_stream
. . . . . . . . > JOIN link_percents AS lp
. . . . . . . . >   ON request = lp.page_title
. . . . . . . . > WHERE prev = 'Electric_Light_Orchestra'
. . . . . . . . > ORDER BY lp.percentage DESC
. . . . . . . . > LIMIT 20;
```

Step 9: Electric_Light_Orchestra


Hotel_California
Eagles_(band)
Eagles_discography
Olivia_Newton-John_albums_discography
Olivia_Newton-John_singles_discography
Xanadu_(soundtrack)
A_Box_of_Their_Best
Four_Light_Years
Electric_Light_Orchestra
....

Final Path

Question 4:

Find an example of an English wikipedia article that is relatively more popular in the UK.
Find the same for the US and Australia.

Methodology

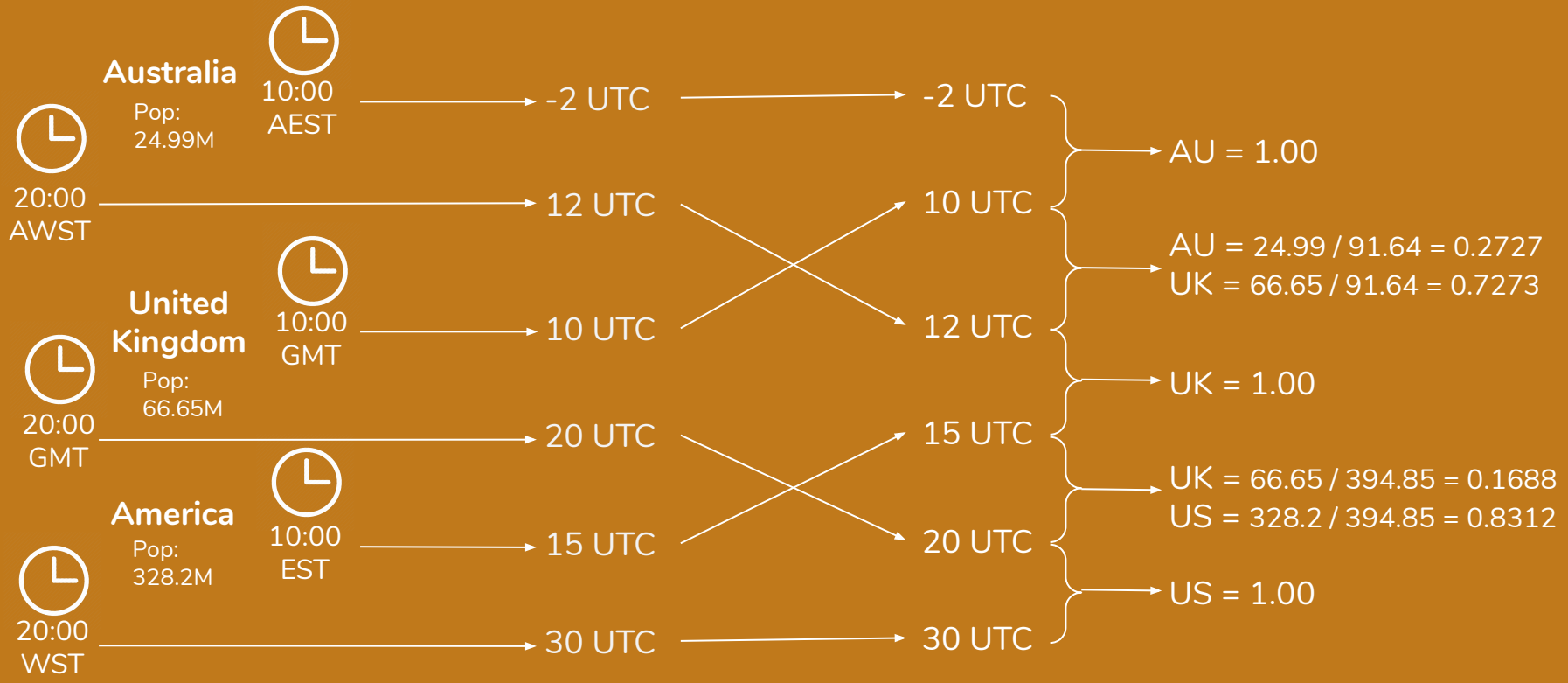
1. Determine appropriate internet activity times for each country.
2. Translate those times into UTC.
3. Determine how to split traffic in overlapping time spans.
4. Create temporary tables for each country to hold data for given time spans. Include traffic split weights.
5. Multiply logged pageviews by weighted fractions and compare.

Assumptions

- People are most likely not using Wikipedia first thing in the morning or last thing at night → 10:00 AM - 8:00 PM (10:00 - 20:00).

- In time periods where there is an overlap between the countries: split traffic based on ratio of populations.

- For the US and Australia (which span multiple time zones) “start” traffic at 10:00 AM in the earliest time zone and end it at 20:00 (8:00 PM) in the latest time zone.



Time zone conversion into timeline of 10/20/2020 in UTC.

```
0: jdbc:hive2://> SELECT *
. . . . . . . . > FROM comparison
. . . . . . . . > WHERE uk_views > us_views
. . . . . . . . >   AND uk_views > au_views
. . . . . . . . > ORDER BY uk_views DESC
. . . . . . . . > LIMIT 20;
```

First SQL Statement: Popular in the UK

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.11 sec   HDFS Read: 209910215 HDFS Write: 1023 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 110 msec
OK
```

comparison.page_title	comparison.us_views	comparison.uk_views	comparison.au_views
F5_Networks	19196	43185	9930
Big_Muskie	10629	20155	18711
List_of_countries_by_GDP_(nominal)	10588	15827	11298
Daniel_Sams_(cricketer)	11364	15614	96
Centenarian	9739	15442	8153
Frankenstein_Castle	1696	8899	5240
Firass_Dirani	2265	8796	3607
Law_&_Order:_Special_Victims_Unit/Season_20	0	6890	2183
American_comic_book	1081	5489	2821
Aaron_Swartz	3471	5351	1927
Rampage_(2018_film)	3143	5320	3788
Michael_Anton	3216	4739	810
Blue_Zone	2754	4035	1725
Postcolonialism	745	4027	1909
Michael_Hastings_(journalist)	2798	3740	221
The_Girl_Next_Door_(2004_film)	1944	3441	1796
Kristian_Opseth	1316	3241	882
Paperback_Hero_(1999_film)	410	2854	1462
cutie_pie	1188	2846	376
Zangilan_District	2385	2437	520

```
20 rows selected (16.678 seconds)
```

First Results: More popular in the UK

```
0: jdbc:hive2://> SELECT *
. . . . . . . . > FROM comparison
. . . . . . . . > WHERE us_views > uk_views
. . . . . . . . >   AND us_views > au_views
. . . . . . . . > ORDER BY us_views DESC
. . . . . . . . > LIMIT 20;
```

Second SQL Statement: Popular in the US

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.57 sec   HDFS Read: 209910215 HDFS Write: 1009 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 570 msec
OK
```

comparison.page_title	comparison.us_views	comparison.uk_views	comparison.au_views
Main_Page	3798632	1364056	2824281
Special:Search	978073	370043	709756
-	345875	131167	252612
Sokushinbutsu	299177	481	465
C._Rajagopalachari	182501	35524	406
The_Haunting_of_Bly_Manor	127259	29919	110074
Mookie_Betts	112607	2078	7416
Three_Red_Banners	108174	2581	152
Chicago_Seven	100049	22759	95529
Bible	98711	33952	74129
Alexandria_Ocasio-Cortez	83299	4344	15339
The_Trial_of_the_Chicago_7	77997	19489	68396
Deaths_in_2020	76690	24740	52651
Abbie_Hoffman	74363	14622	73072
Harshad_Mehta	71122	29513	45138
2016_United_States_presidential_election	66849	15803	47749
Hunter_Biden	64820	15669	52276
Joe_Biden	63821	16946	50360
End_SARS	63655	6691	11496
Tyler_Glasnow	63459	758	1975

```
20 rows selected (16.981 seconds)
```

Second Results: More popular in the US

```
0: jdbc:hive2://> SELECT *
. . . . . . . . > FROM comparison
. . . . . . . . > WHERE au_views > uk_views
. . . . . . . . >   AND au_views > us_views
. . . . . . . . > ORDER BY au_views DESC
. . . . . . . . > LIMIT 20;
```

Third SQL Statement: Popular in Australia

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.84 sec   HDFS Read: 209910215 HDFS Write: 1040 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 840 msec
OK
```

comparison.page_title	comparison.us_views	comparison.uk_views	comparison.au_views
Jeffrey_Toobin	168187	57121	193636
Kyler_Murray	13551	7685	97635
Robert_Redford	70816	41428	97483
Sisters_at_Heart	3690	4725	97258
Jeff_Bridges	59613	35354	84707
Dancing_with_the_Stars_(American_season_29)	20922	8159	79787
Murder_of_Robert_McCartney	2476	24092	78257
Killing_in_the_Name	914	1064	51174
Microsoft_Office	47060	19352	49238
SSC_Tuatara	39627	17659	47022
Andy_Dalton	6318	3007	47019
Lovecraft_Country_(TV_series)	39734	8472	40819
Anthony_Fauci	26076	10437	40639
Josh_Allen_(quarterback)	3655	1573	34549
Patrick_Mahomes	8180	2857	34108
Watts_family_murders	32782	9283	32817
Lola_Van_Wagenen	25165	14788	31893
Chrishell_Stause	11890	2522	31754
Kirstie_Alley	11169	5777	31050
Budda_Baker	6633	2626	30667

```
20 rows selected (16.006 seconds)
```

Third Results: More popular in Australia

Question 5:

How many users will see the average vandalized wikipedia page before the offending edit is reversed?

Methodology

1. Collect pertinent subset of data from the massive revision data file (i.e. page_title, revision_seconds_to_identity_revert) and the total pageviews aggregation.

2. Transform the gathered data into necessary data format: average revert time in minutes, average number of pageviews per minute.

Assumptions

- Average view per page per minute is calculated by summing total_views of all pages in enwiki and dividing by the number of minutes represented (i.e. 32 hours = 1920 minutes).

- In a more thorough analysis, we would match up page_title from the pageviews data and the page_title from our revision data for a better data match.

```
0: jdbc:hive2://> SELECT Round(AVG(revert_time_min),2) AS avg_revert_time_min
. . . . . > FROM vandalism;
```

```
0: jdbc:hive2://> SELECT count(page_title) AS total_pages
. . . . . > FROM total_views;
```

```
0: jdbc:hive2://> SELECT sum(total_views) AS gross_views
. . . . . > FROM total_views;
```

avg_revert_time_min
1696.98

total_pages
6303194

gross_views
341305872

total_time
32 hours x 60 min =
1920

[(gross_views/total_pages) / mins in dataset] x avg_revert_time_min

[(341305872 / 6303194) / 1920] * 1696.98 =

~48 views before reversion.

Data Calculations

Question 6:

Run an analysis you find interesting on the wikipedia datasets we're using.

My Selection:

Which of my favorite 10 animes has the highest view total?

Methodology

Similar to question one, I'll check and compare the total pageviews of 10 of my favorite anime shows, which are:

1. Attack on Titan
2. Baccano!
3. Cowboy Bebop
4. Crayon Shin-chan
5. Death Note
6. Eden of the East
7. FLCL
8. Ghost in the Shell
9. Mushishi
10. Spice and Wolf

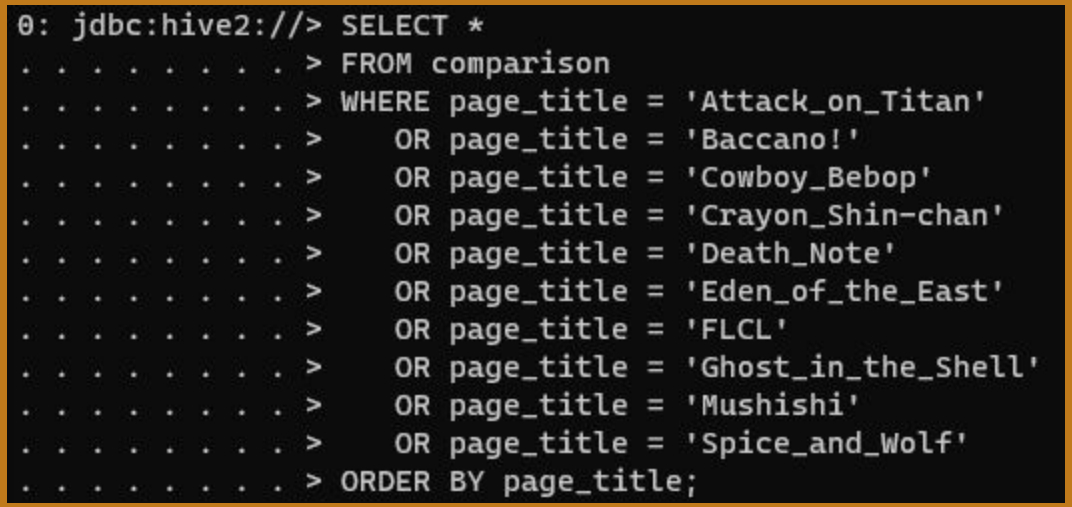
```
0: jdbc:hive2://> SELECT *
. . . . . > FROM total_views
. . . . . > WHERE page_title = 'Attack_on_Titan'
. . . . . >    OR page_title = 'Baccano!'
. . . . . >    OR page_title = 'Cowboy_Bebop'
. . . . . >    OR page_title = 'Crayon_Shin-chan'
. . . . . >    OR page_title = 'Death_Note'
. . . . . >    OR page_title = 'Eden_of_the_East'
. . . . . >    OR page_title = 'FLCL'
. . . . . >    OR page_title = 'Ghost_in_the_Shell'
. . . . . >    OR page_title = 'Mushishi'
. . . . . >    OR page_title = 'Spice_and_Wolf'
. . . . . > ORDER BY total_views DESC;
```

Compound SQL Query

total_views.page_title	total_views.total_views
Attack_on_Titan	12439
Death_Note	7095
Cowboy_Bebop	4706
Ghost_in_the_Shell	2781
Crayon_Shin-chan	2553
FLCL	1856
Spice_and_Wolf	753
Baccano!	684
Mushishi	680
Eden_of_the_East	439

10 rows selected (15.541 seconds)

Results



Compound SQL Query

comparison.page_title	comparison.us_views	comparison.uk_views	comparison.au_views
Attack_on_Titan	6079	1809	4551
Baccano!	317	87	280
Cowboy_Bebop	2339	515	1852
Crayon_Shin-chan	1071	517	965
Death_Note	3321	1047	2727
Eden_of_the_East	208	59	172
FLCL	857	214	785
Ghost_in_the_Shell	1401	288	1092
Mushishi	319	85	276
Spice_and_Wolf	366	89	298

Results

Github Repository

https://github.com/sean-horner/revature_project_1.git

Questions?