Project 1: Wikipedia Data Analysis Project

ROGER GRIFFIN

Question 1:
Which English Wikipedia article got the most traffic on October 20?

- To answer, first download data for all of October 20th. For this project, October 20th, 2020 was used.
- With the downloaded information, run a Hadoop MapReduce to condense data and filter the information to only English Wikipedia articles.
- This data was loaded into a table in Hive for easy querying.

- Other than which October 20th was meant used for this project, no assumptions had to be made.

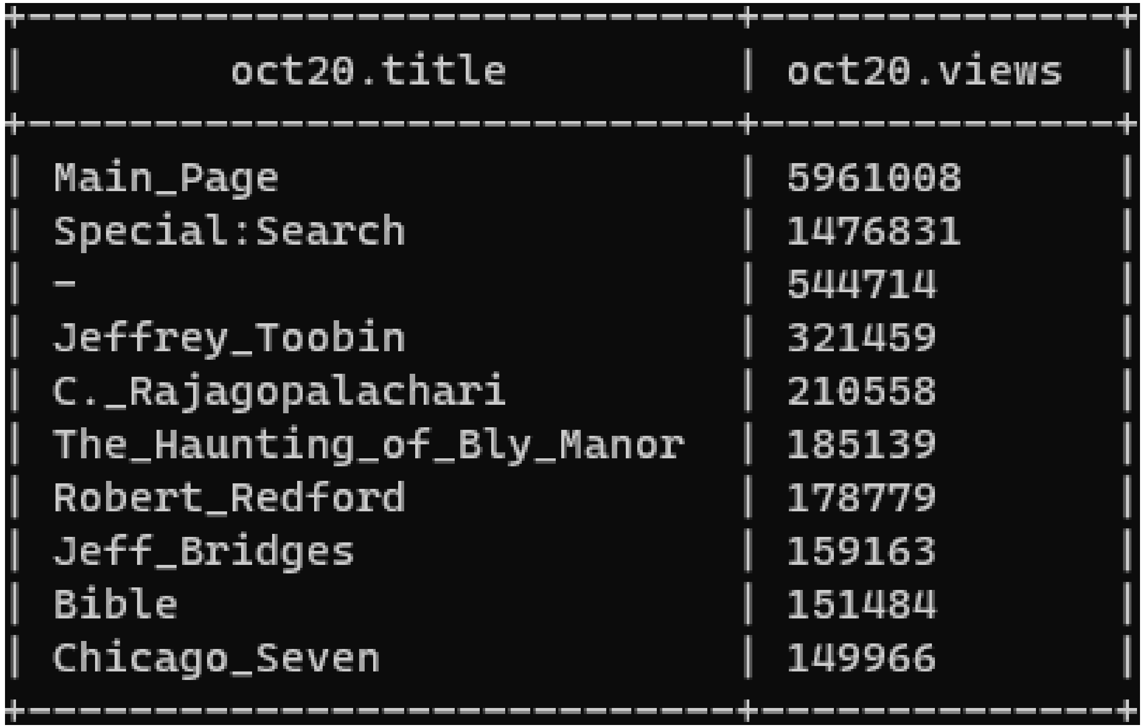
Question 1(cont)

- SQL Query:

SELECT *

FROM OCT20

ORDER BY VIEWS DESC LIMIT 10;



```
+-----------------------------+---------------+
|         oct20.title         |  oct20.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        |
+-----------------------------+---------------+
```

Question 2: What English Wikipedia article has the largest fraction of its readers follow an internal link to another Wikipedia article?

- To answer this, the clickstream dump for September of 2020 was used.
- Clickstream data is formatted:

Previous Page	Current Page	Type	Number
Where the user came from	Page currently on	Did they follow another Wikipedia page link/did they come from an external link	Number of users that got to this page that way

- Knowing the format and having all the data needed, I ran 2 map reduces on the same data. They did:
  1. A MapReduce that made the key the current page's title and combined the views data, getting total views of that page
  2. A MapReduce that looked for the third part of the entry to be "link" and combined those using Previous page as the key, getting how many links were clicked from that page.
- These reduced data sets were then loaded into separate hive tables and queried together. Eventually made into a single table to make the query easier
- The data was queried to include a statement only allowing for articles with over a million views as to give more accurate data. Under a million views ran into a problem where you would get more links pressed than actual views on a page

Question 2(cont)

- SQL Query:

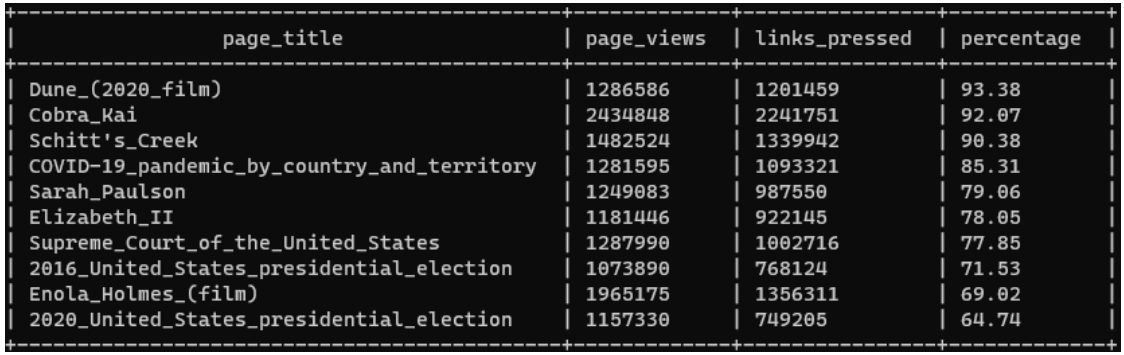
SELECT PAGE_TITLE, PAGE_VIEWS, LINKS_PRESSED, ROUND((LINKS_PRESSED/PAGE_VIEWS)*100, 2) AS PERCENTAGE

FROM CLICKSTREAMFINAL

WHERE PAGE_VIEWS > 1000000 --limits data to relevant data

ORDER BY PERCENTAGE DESC

LIMIT 10;



```
+--------------------------------------------+------------+---------------+------------+
|                 page_title                 | page_views | links_pressed | percentage |
+--------------------------------------------+------------+---------------+------------+
| Dune_(2020_film)                           | 1286586    | 1201459       | 93.38      |
| Cobra_Kai                                  | 2434848    | 2241751       | 92.07      |
| Schitt's_Creek                             | 1482524    | 1339942       | 90.38      |
| COVID-19_pandemic_by_country_and_territory | 1281595    | 1093321       | 85.31      |
| Sarah_Paulson                              | 1249083    | 987550        | 79.06      |
| Elizabeth_II                               | 1181446    | 922145        | 78.05      |
| Supreme_Court_of_the_United_States         | 1287990    | 1002716       | 77.85      |
| 2016_United_States_presidential_election   | 1073890    | 768124        | 71.53      |
| Enola_Holmes_(film)                        | 1965175    | 1356311       | 69.02      |
| 2020_United_States_presidential_election   | 1157330    | 749205        | 64.74      |
+--------------------------------------------+------------+---------------+------------+
```

Question 3: What series of Wikipedia articles, starting with [Hotel California], keeps the largest fraction of its readers clicking on internal links?

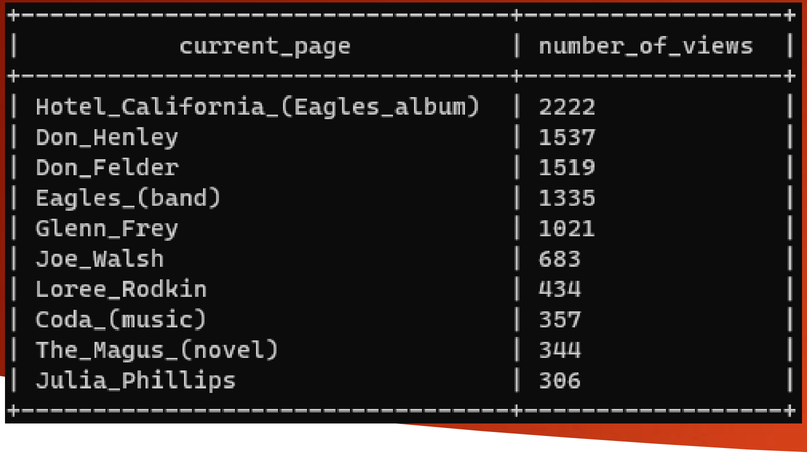
- This required the complete Clickstream used in Question 2, so data was already downloaded
  - Since format is already known, can be loaded into the table with names that mean more to the end user.
- Loaded complete Clickstream into Hive Database
- Did multiple of the same query to get final result

Question 3(cont.)

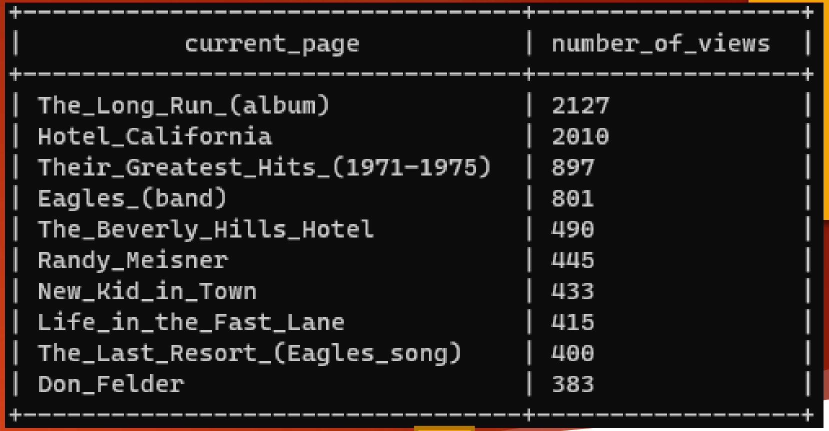
- SQL Query:

```
SELECT CURRENT_PAGE, NUMBER_OF_VIEWS
FROM CLICKSTREAMUNEDIT
WHERE PREVIOUS_PAGE LIKE “insert previous title here" AND LINK_TYPE = "link"
AND NOT(PREVIOUS_PAGE = "other-internal" OR PREVIOUS_PAGE="other-search" OR
PREVIOUS_PAGE="other-external" OR PREVIOUS_PAGE="other-empty" OR
PREVIOUS_PAGE="other-other")
ORDER BY NUMBER_OF_VIEWS DESC
limit 10;
```

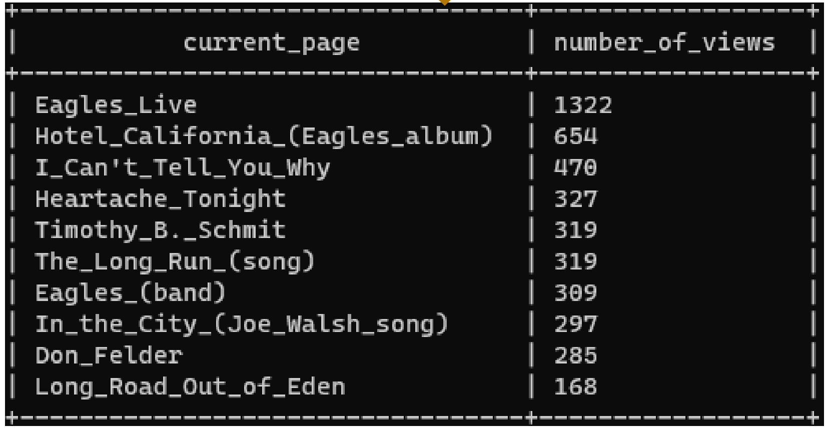
- Hotel_California -> Hotel_California_(Eagles_album) (2222) -> The_Long_Run_(album) (2127) -> Eagles_Live (1322) -> Eagles_Greatest_Hits,_Vol._2 (1136) ->The_Very_Best_of_the_Eagles (996)



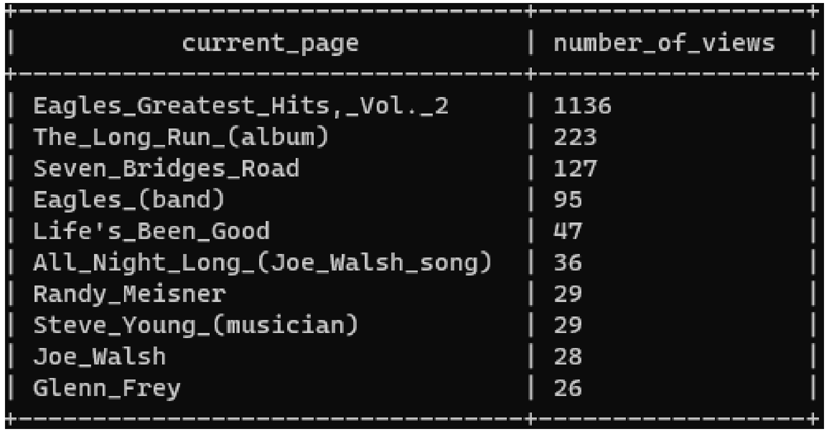
```
+---------------------------------+-----------------+
|          current_page           | number_of_views |
+---------------------------------+-----------------+
| Hotel_California_(Eagles_album) | 2222            |
| Don_Henley                      | 1537            |
| Don_Felder                      | 1519            |
| Eagles_(band)                   | 1335            |
| Glenn_Frey                      | 1021            |
| Joe_Walsh                       | 683             |
| Loree_Rodkin                    | 434             |
| Coda_(music)                    | 357             |
| The_Magus_(novel)               | 344             |
| Julia_Phillips                  | 306             |
+---------------------------------+-----------------+
```



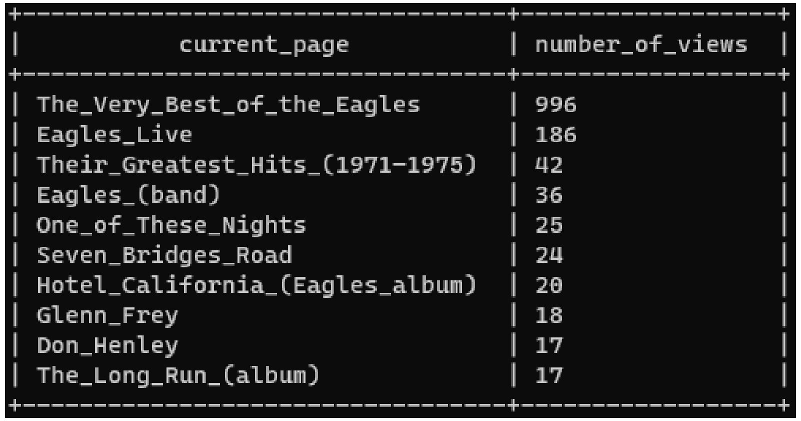
```
+---------------------------------+-----------------+
|          current_page           | number_of_views |
+---------------------------------+-----------------+
| The_Long_Run_(album)            | 2127            |
| Hotel_California                | 2010            |
| Their_Greatest_Hits_(1971-1975) | 897             |
| Eagles_(band)                   | 801             |
| The_Beverly_Hills_Hotel         | 490             |
| Randy_Meisner                   | 445             |
| New_Kid_in_Town                 | 433             |
| Life_in_the_Fast_Lane           | 415             |
| The_Last_Resort_(Eagles_song)   | 400             |
| Don_Felder                      | 383             |
+---------------------------------+-----------------+
```



```
+---------------------------------+-----------------+
|          current_page           | number_of_views |
+---------------------------------+-----------------+
| Eagles_Live                     | 1322            |
| Hotel_California_(Eagles_album) | 654             |
| I_Can't_Tell_You_Why            | 470             |
| Heartache_Tonight               | 327             |
| Timothy_B._Schmit               | 319             |
| The_Long_Run_(song)             | 319             |
| Eagles_(band)                   | 309             |
| In_the_City_(Joe_Walsh_song)    | 297             |
| Don_Felder                      | 285             |
| Long_Road_Out_of_Eden           | 168             |
+---------------------------------+-----------------+
```



```
+---------------------------------+-----------------+
|          current_page           | number_of_views |
+---------------------------------+-----------------+
| Eagles_Greatest_Hits,_Vol._2    | 1136            |
| The_Long_Run_(album)            | 223             |
| Seven_Bridges_Road              | 127             |
| Eagles_(band)                   | 95              |
| Life's_Been_Good                | 47              |
| All_Night_Long_(Joe_Walsh_song) | 36              |
| Randy_Meisner                   | 29              |
| Steve_Young_(musician)          | 29              |
| Joe_Walsh                       | 28              |
| Glenn_Frey                      | 26              |
+---------------------------------+-----------------+
```



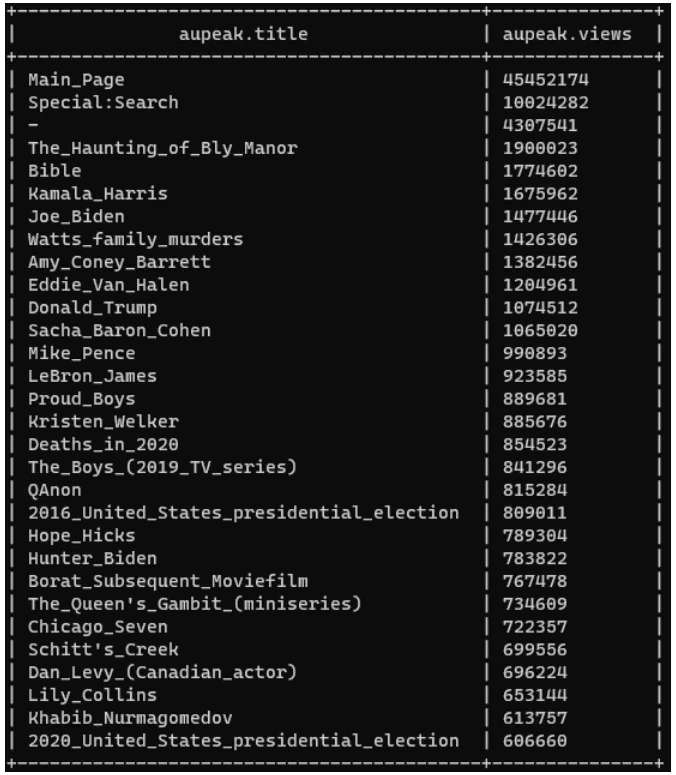
```
+---------------------------------+-----------------+
|          current_page           | number_of_views |
+---------------------------------+-----------------+
| The_Very_Best_of_the_Eagles     | 996             |
| Eagles_Live                     | 186             |
| Their_Greatest_Hits_(1971-1975) | 42              |
| Eagles_(band)                   | 36              |
| One_of_These_Nights             | 25              |
| Seven_Bridges_Road              | 24              |
| Hotel_California_(Eagles_album) | 20              |
| Glenn_Frey                      | 18              |
| Don_Henley                      | 17              |
| The_Long_Run_(album)            | 17              |
+---------------------------------+-----------------+
```

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.

- This was done on the data from October of 2020, and was done similar to Question 1.
  - Instead of doing 24 hours of 1 day, it was done for limited hours each day over 30 days.
  - This was done 3 times for the peak times for each country, considering that the peak hours include their specific 5PM-9PM times.
    - For the US and Australia, it was 5PM in the Easternmost time-zone to 9PM in the Westernmost time-zone.
- The results of these 3 MapReduces are loaded into different tables for separate querying.

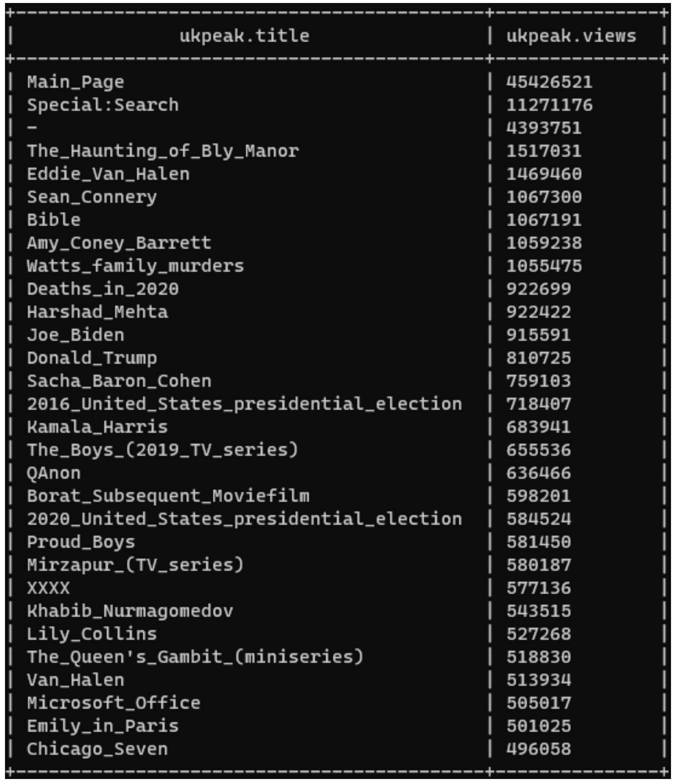
Question 4(cont)

Australia Peak Query:
SELECT * FROM AUPEAK
ORDER BY VIEWS DESC
LIMIT 10;



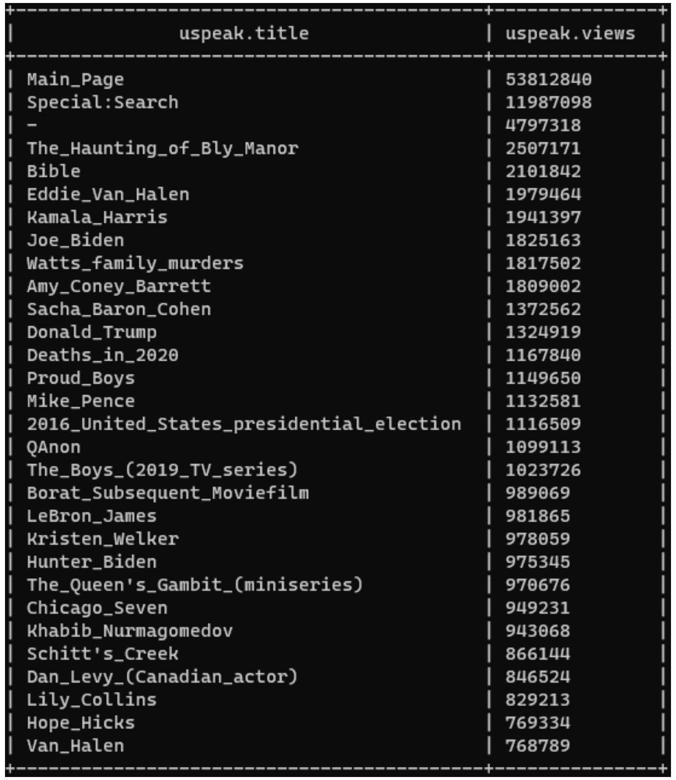
```
+-------------------------------------------+-----------------+
|               aupeak.title                | aupeak.views    |
+-------------------------------------------+-----------------+
| Main_Page                                 | 45452174        |
| Special:Search                            | 10024282        |
| -                                         | 4307541         |
| The_Haunting_of_Bly_Manor                 | 1900023         |
| Bible                                     | 1774602         |
| Kamala_Harris                             | 1675962         |
| Joe_Biden                                 | 1477446         |
| Watts_family_murders                      | 1426306         |
| Amy_Coney_Barrett                         | 1382456         |
| Eddie_Van_Halen                           | 1204961         |
| Donald_Trump                              | 1074512         |
| Sacha_Baron_Cohen                         | 1065020         |
| Mike_Pence                                | 990893          |
| LeBron_James                              | 923585          |
| Proud_Boys                                | 889681          |
| Kristen_Welker                            | 885676          |
| Deaths_in_2020                            | 854523          |
| The_Boys_(2019_TV_series)                 | 841296          |
| QAnon                                     | 815284          |
| 2016_United_States_presidential_election  | 809011          |
| Hope_Hicks                                | 789304          |
| Hunter_Biden                              | 783822          |
| Borat_Subsequent_Moviefilm                | 767478          |
| The_Queen's_Gambit_(miniseries)           | 734609          |
| Chicago_Seven                             | 722357          |
| Schitt's_Creek                            | 699556          |
| Dan_Levy_(Canadian_actor)                 | 696224          |
| Lily_Collins                              | 653144          |
| Khabib_Nurmagomedov                       | 613757          |
| 2020_United_States_presidential_election  | 606660          |
+-------------------------------------------+-----------------+
```

UK Peak Query:
SELECT * FROM UKPEAK
ORDER BY VIEWS DESC
LIMIT 10;



```
+-------------------------------------------+-----------------+
|               ukpeak.title                | ukpeak.views    |
+-------------------------------------------+-----------------+
| Main_Page                                 | 45426521        |
| Special:Search                            | 11271176        |
| -                                         | 4393751         |
| The_Haunting_of_Bly_Manor                 | 1517031         |
| Eddie_Van_Halen                           | 1469460         |
| Sean_Connery                              | 1067300         |
| Bible                                     | 1067191         |
| Amy_Coney_Barrett                         | 1059238         |
| Watts_family_murders                      | 1055475         |
| Deaths_in_2020                            | 922699          |
| Harshad_Mehta                             | 922422          |
| Joe_Biden                                 | 915591          |
| Donald_Trump                              | 810725          |
| Sacha_Baron_Cohen                         | 759103          |
| 2016_United_States_presidential_election  | 718407          |
| Kamala_Harris                             | 683941          |
| The_Boys_(2019_TV_series)                 | 655536          |
| QAnon                                     | 636466          |
| Borat_Subsequent_Moviefilm                | 598201          |
| 2020_United_States_presidential_election  | 584524          |
| Proud_Boys                                | 581450          |
| Mirzapur_(TV_series)                      | 580187          |
| XXXX                                      | 577136          |
| Khabib_Nurmagomedov                       | 543515          |
| Lily_Collins                              | 527268          |
| The_Queen's_Gambit_(miniseries)           | 518830          |
| Van_Halen                                 | 513934          |
| Microsoft_Office                          | 505017          |
| Emily_in_Paris                            | 501025          |
| Chicago_Seven                             | 496058          |
+-------------------------------------------+-----------------+
```

US Peak Query:
SELECT * FROM USPEAK
ORDER BY VIEWS DESC
LIMIT 10;



```
+-------------------------------------------+-----------------+
|               uspeak.title                | uspeak.views    |
+-------------------------------------------+-----------------+
| Main_Page                                 | 53812840        |
| Special:Search                            | 11987098        |
| -                                         | 4797318         |
| The_Haunting_of_Bly_Manor                 | 2507171         |
| Bible                                     | 2101842         |
| Eddie_Van_Halen                           | 1979464         |
| Kamala_Harris                             | 1941397         |
| Joe_Biden                                 | 1825163         |
| Watts_family_murders                      | 1817502         |
| Amy_Coney_Barrett                         | 1809002         |
| Sacha_Baron_Cohen                         | 1372562         |
| Donald_Trump                              | 1324919         |
| Deaths_in_2020                            | 1167840         |
| Proud_Boys                                | 1149650         |
| Mike_Pence                                | 1132581         |
| 2016_United_States_presidential_election  | 1116509         |
| QAnon                                     | 1099113         |
| The_Boys_(2019_TV_series)                 | 1023726         |
| Borat_Subsequent_Moviefilm                | 989069          |
| LeBron_James                              | 981865          |
| Kristen_Welker                            | 978059          |
| Hunter_Biden                              | 975345          |
| The_Queen's_Gambit_(miniseries)           | 970676          |
| Chicago_Seven                             | 949231          |
| Khabib_Nurmagomedov                       | 943068          |
| Schitt's_Creek                            | 866144          |
| Dan_Levy_(Canadian_actor)                 | 846524          |
| Lily_Collins                              | 829213          |
| Hope_Hicks                                | 769334          |
| Van_Halen                                 | 768789          |
+-------------------------------------------+-----------------+
```

Question 4(cont)

- Analyzing the data we can see duplicates between different countries, so those wouldn't be considered unique.
  - With duplicate information removed, we can see things more unique to specific countries.

1. Australia: Dan Levy, Hope Hicks
2. UK: Sean Connery, Emily In Paris, XXXX
3. US: Kristen Welker

Question 5: Analyze how many users will see the average vandalized Wikipedia page before the offending edit is reversed.

- For this, we need the Page Revisions and User History dump. We can use the October of 2020 dump.
- This data has 70 different columns of potential information, putting that directly into a hive database takes time, but makes writing searches easier on the programmer.
- With the data loaded, we can find an average time to revise an edit back to what it was prior to the edit by averaging the revision_seconds_to_identify_revert.
  - Using a where statement we can limit this to edits over an hour. This assumption is used to try to remove reverts by people who posted something accidentally and removed it themselves.
- We can then use the data from Question 1 to find the average page views per day(assuming October 20th was an average day), divide that number to get the average hourly views on a page and find out how many people could see the offending page.

Question 5(cont)

- Revision Query:

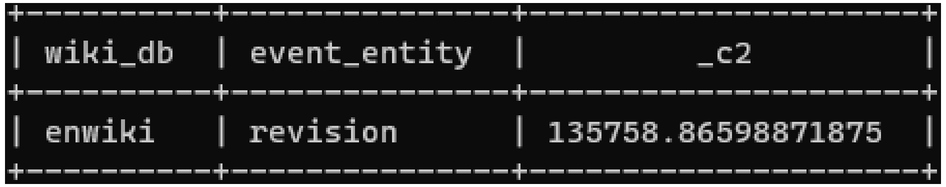
SELECT WIKI_DB, EVENT_ENTITY, AVG(REVISION_SECONDS_TO_IDENTITY_REVERT)

FROM REVISIONS

WHERE REVISION_SECONDS_TO_IDENTITY_REVERT > 600

GROUP BY WIKI_DB, EVENT_ENTITY

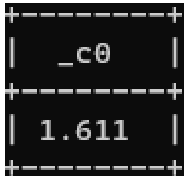
LIMIT 10;



```
+----------+----------------+----------------------+
| wiki_db  | event_entity   |                  _c2 |
+----------+----------------+----------------------+
| enwiki   | revision       | 135758.86598871875   |
+----------+----------------+----------------------+
```

- Average Hourly Query:

SELECT ROUND(AVG(VIEWS)/24, 3) FROM OCT20;



```
+--------+
|  _c0   |
+--------+
| 1.611  |
+--------+
```

- This result with an average of pageviews per hour gives a result of:

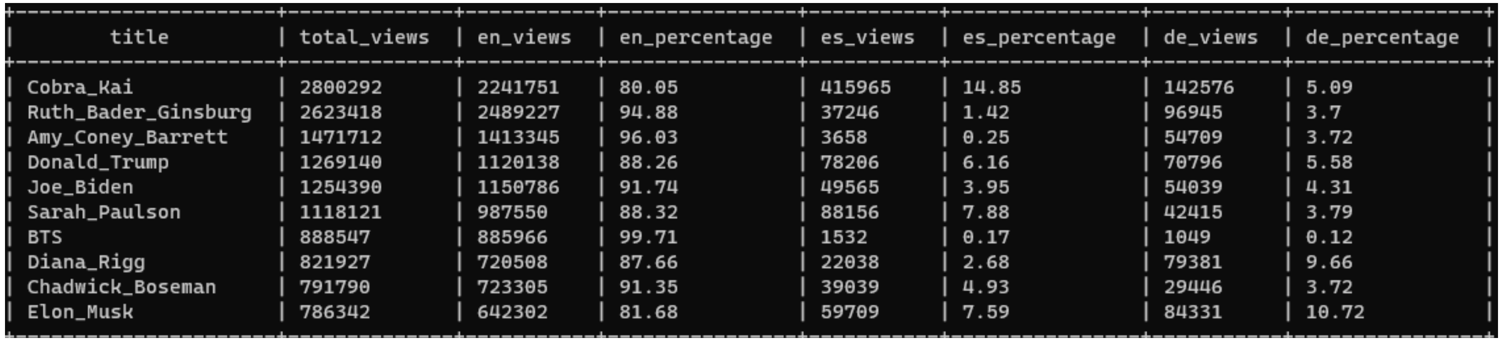
37.7 hours * 1.611 views/hour is about 61 views.

Question 6: Find pages that users view on the English, Spanish and German wikipedias and show the percentage of views each site gave.

- Need data from all 3 Wikis, so we'll go back to using the clickstream for an easier time since it's the same process for the MapReduce as Question 2.
- With the total views for each Wikipedia now reduced, we can now load those into individual tables
- With them in individual tables, we can use inner joins and round() functions to calculate what we're looking.

Question 6(cont)

```
SELECT TITLE, TOTAL_VIEWS, EN_VIEWS, ROUND((EN_VIEWS/TOTAL_VIEWS)*100, 2) AS
EN_PERCENTAGE, ES_VIEWS, ROUND((ES_VIEWS/TOTAL_VIEWS)*100, 2) AS ES_PERCENTAGE,
DE_VIEWS, ROUND((DE_VIEWS/TOTAL_VIEWS)*100, 2) AS DE_PERCENTAGE
 FROM MULTILANGCLICKSTREAM
 INNER JOIN ENCLICKSTREAM ON MULTILANGCLICKSTREAM.TITLE=ENCLICKSTREAM.EN_TITLE
INNER JOIN ESCLICKSTREAM ON MULTILANGCLICKSTREAM.TITLE=ESCLICKSTREAM.ES_TITLE
 INNER JOIN DECLICKSTREAM ON MULTILANGCLICKSTREAM.TITLE=DECLICKSTREAM.DE_TITLE
 ORDER BY TOTAL_VIEWS DESC
 LIMIT 10;
```



```
+-----------------------+--------------+-----------+----------------+-----------+-----------------+-----------+----------------+
|         title         | total_views  | en_views  | en_percentage  | es_views  | es_percentage   | de_views  | de_percentage  |
+-----------------------+--------------+-----------+----------------+-----------+-----------------+-----------+----------------+
| Cobra_Kai             | 2800292      | 2241751   | 80.05          | 415965    | 14.85           | 142576    | 5.09           |
| Ruth_Bader_Ginsburg   | 2623418      | 2489227   | 94.88          | 37246     | 1.42            | 96945     | 3.7            |
| Amy_Coney_Barrett     | 1471712      | 1413345   | 96.03          | 3658      | 0.25            | 54709     | 3.72           |
| Donald_Trump          | 1269140      | 1120138   | 88.26          | 78206     | 6.16            | 70796     | 5.58           |
| Joe_Biden             | 1254390      | 1150786   | 91.74          | 49565     | 3.95            | 54039     | 4.31           |
| Sarah_Paulson         | 1118121      | 987550    | 88.32          | 88156     | 7.88            | 42415     | 3.79           |
| BTS                   | 888547       | 885966    | 99.71          | 1532      | 0.17            | 1049      | 0.12           |
| Diana_Rigg            | 821927       | 720508    | 87.66          | 22038     | 2.68            | 79381     | 9.66           |
| Chadwick_Boseman      | 791790       | 723305    | 91.35          | 39039     | 4.93            | 29446     | 3.72           |
| Elon_Musk             | 786342       | 642302    | 81.68          | 59709     | 7.59            | 84331     | 10.72          |
+-----------------------+--------------+-----------+----------------+-----------+-----------------+-----------+----------------+
```

Questions?

Thank you for listening