project one

by j.rice

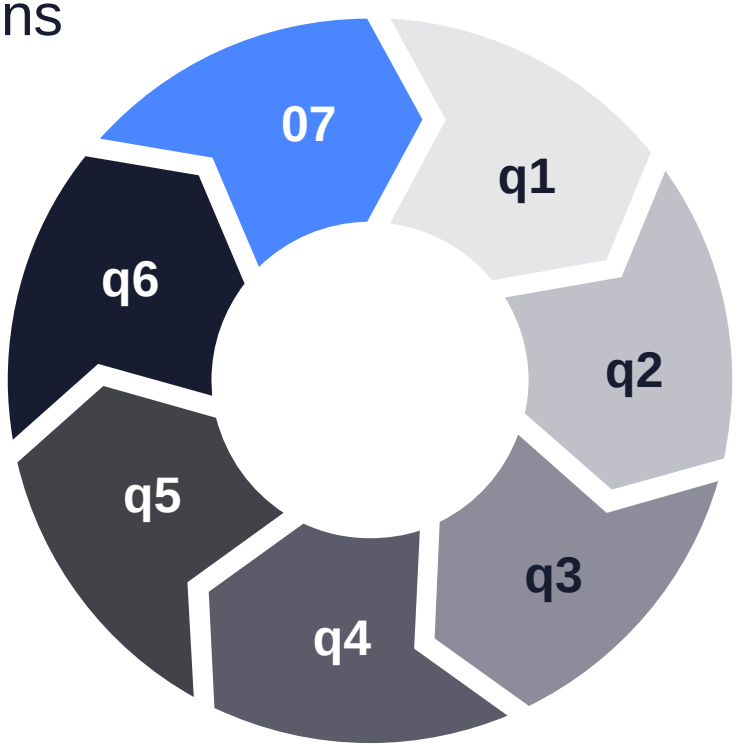
presentation
agenda

07 conclusions

q6 – comparison of mobile vs. desktop views on articles

q5 – average vandalized article visits before reversal

q4 – articles more popular in US, UK & Australia

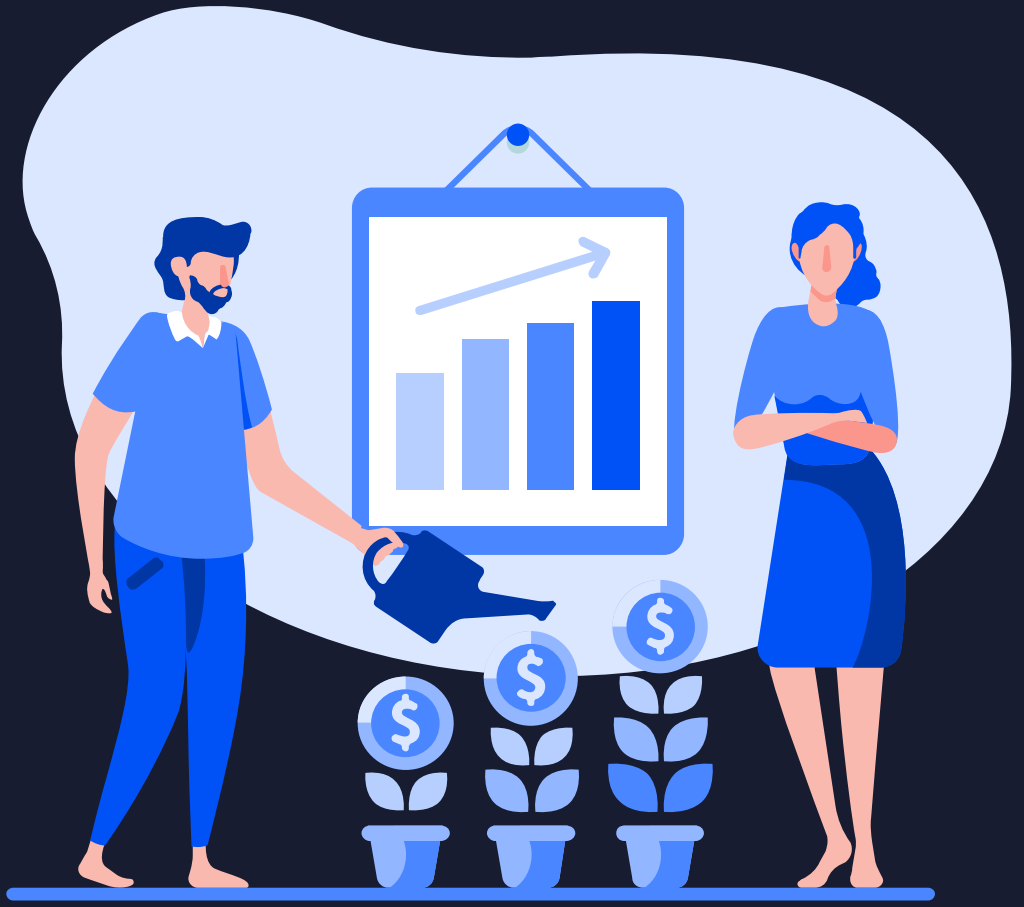


q1 – article with the most traffic on oct 20th

q2 – article with the largest % of internal link visits

q3 – series of Hotel California articles with the highest % internal link visits

what article on Oct 20th got the most traffic?



q1

PROCESS

- Ran a MapReduce to grab on English results
- Limited and sorted by total views
- grouped by article and domain code to show device specifics

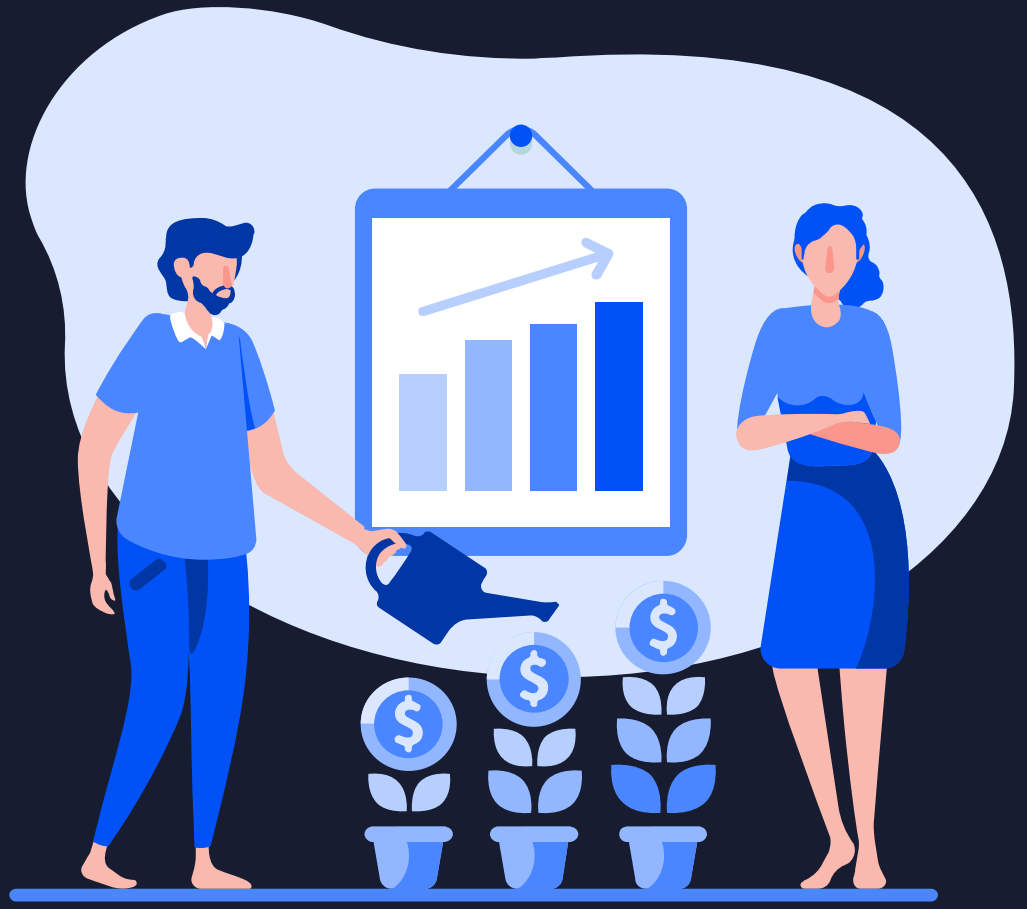
q1

RESULTS

```
Total MapReduce CPU Time Spent: 8 minutes
OK
+----------------------------+----------+
|        article_name        |  total   |
+----------------------------+----------+
| 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 (147.629 seconds)
```

```
Total MapReduce CPU Time Spent: 8 minutes 45 seconds 990
OK
+--------------+----------------------------+----------+
| domain_code  |        article_name        |  total   |
+--------------+----------------------------+----------+
| en.m         | Main_Page                  | 3234621  |
| en           | Main_Page                  | 2726387  |
| en           | Special:Search             | 910309   |
| en.m         | Special:Search             | 566522   |
| en.m         | -                          | 419824   |
| en.m         | Jeffrey_Toobin             | 204735   |
| en.m         | C._Rajagopalachari         | 199383   |
| en           | Bible                      | 148726   |
| en.m         | The_Haunting_of_Bly_Manor  | 147681   |
| en           | -                          | 124890   |
+--------------+----------------------------+----------+
10 rows selected (153.375 seconds)
```

what article has the largest % of internal link visits?



q2

PROCESS

- Hive/Hadoop Map Reduce on Clickstream & Pageview Data
- Joined both Clickstream & Pageview tables in Hive
- Ran a query to divide internal clicks by the total page views

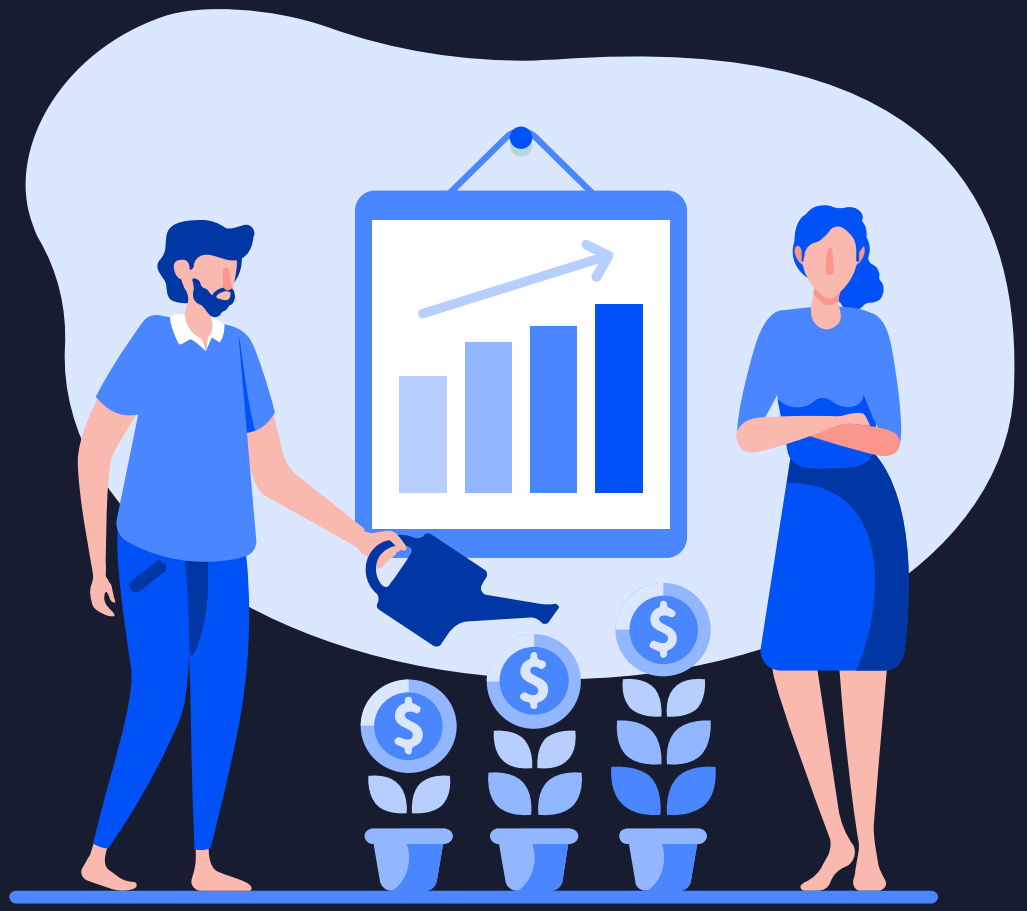
q2

q2_results.article	q2_results.total_views	q2_results.internal_link	q2_results.percent_internal
Main_Page	165044119	2379287	1.442
Ruth_Bader_Ginsburg	7605356	2489227	32.730
Amy_Coney_Barrett	5924508	1413345	23.856
Tenet_(film)	3877047	1386086	35.751
Shooting_of_Breonna_Taylor	3850524	247198	6.420
Dennis_Nilsen	3564441	660393	18.527
Deaths_in_2020	3316200	1595715	48.119
Mulan_(2020_film)	3239724	1749519	54.002
The_Boys_(2019_TV_series)	3184665	2006351	63.000
Bible	3170711	32110	1.013

q2_results.article	q2_results.total_views	q2_results.internal_link	q2_results.percent_internal
Cobra_Kai	2459988	2241751	91.129
Enola_Holmes_(film)	1980000	1356311	68.501
Ratched_(TV_series)	2626716	1668477	63.520
The_Boys_(2019_TV_series)	3184665	2006351	63.000
Donald_Trump	1830929	1120138	61.179
The_Devil_All_the_Time_(film)	1886635	1071565	56.798
Mulan_(2020_film)	3239724	1749519	54.002
Deaths_in_2020	3316200	1595715	48.119
Joe_Biden	2740959	1150786	41.985
September_11_attacks	2028774	850181	41.906

10 rows selected (0.075 seconds)

what series of Hotel California-related articles have the largest amount of internal link visits?



q3

PROCESS

- Queried to select internal links from where the referrer was an article, while the requested title was set to Hotel California
- Queried to select internal links from where the referrer was Hotel California, while the requested title would be any title
- From there, I replaced the where statement portion where “referrer like” “article” was replaced with the next highest article in respect to internal links.

q3

RESULTS

```
0: jdbc:hive2://> select * from link_trace_cs
. . . . . . . . >     where type_traffic="link" and  referrer like "Hotel_California" and not (referrer = "other-internal" or referrer= "other-search" or
. . . . . . . . >     referrer="other-external"or referrer="other-empty"or referrer ="other-other")
. . . . . . . . >     sort by occurences desc
. . . . . . . . >     limit 100;
```

```
Total MapReduce CPU Time Spent: 27 seconds 460 msec
OK
+--------------------------+-------------------------------------------------+-----------------------------+---------------------------+
| link_trace_cs.referrer   |        link_trace_cs.requested_article          | link_trace_cs.type_traffic  | link_trace_cs.occurences  |
+--------------------------+-------------------------------------------------+-----------------------------+---------------------------+
| Hotel_California         | Hotel_California_(Eagles_album)                 | link                        | 2222                      |
| Hotel_California         | Don_Henley                                      | link                        | 1537                      |
| Hotel_California         | Don_Felder                                      | link                        | 1519                      |
| Hotel_California         | Eagles_(band)                                   | link                        | 1335                      |
| Hotel_California         | Glenn_Frey                                      | link                        | 1021                      |
| Hotel_California         | Joe_Walsh                                       | link                        | 683                       |
| Hotel_California         | Loree_Rodkin                                    | link                        | 434                       |
| Hotel_California         | Coda_(music)                                    | link                        | 357                       |
| Hotel_California         | The_Magus_(novel)                               | link                        | 344                       |
| Hotel_California         | Julia_Phillips                                  | link                        | 306                       |
| Hotel_California         | The_Beverly_Hills_Hotel                         | link                        | 297                       |
| Hotel_California         | Life_in_the_Fast_Lane                           | link                        | 286                       |
```

q3

RESULTS

```
. . . . . >     where type_traffic="link" and referrer like "Hotel_California_(Eagles_album)" and not (referrer = "other-internal" or referrer
= "other-search" or
. . . . . >     referrer="other-external"or referrer="other-empty"or referrer ="other-other")
. . . . . >     sort by occurences desc
. . . . . >     limit 100;
```

link_trace_cs.referrer	link_trace_cs.requested_article	link_trace_cs.type_traffic	link_trace_cs.occurences
Hotel_California_(Eagles_album)	The_Long_Run_(album)	link	2127
Hotel_California_(Eagles_album)	Hotel_California	link	2010
Hotel_California_(Eagles_album)	Their_Greatest_Hits_(1971-1975)	link	897
Hotel_California_(Eagles_album)	Eagles_(band)	link	801
Hotel_California_(Eagles_album)	The_Beverly_Hills_Hotel	link	490
Hotel_California_(Eagles_album)	Randy_Meisner	link	445
Hotel_California_(Eagles_album)	New_Kid_in_Town	link	433
Hotel_California_(Eagles_album)	Life_in_the_Fast_Lane	link	415
Hotel_California_(Eagles_album)	The_Last_Resort_(Eagles_song)	link	400

q3

RESULTS

```
0: jdbc:hive2://> select * from link_trace_cs
. . . . . . . . >     where type_traffic="link" and referrer like "The_Long_Run_(album)" and not (referrer = "other-internal" or referrer= "other-se
arch" or
. . . . . . . . >     referrer="other-external"or referrer="other-empty"or referrer ="other-other")
. . . . . . . . >     sort by occurences desc
. . . . . . . . >     limit 100;
```

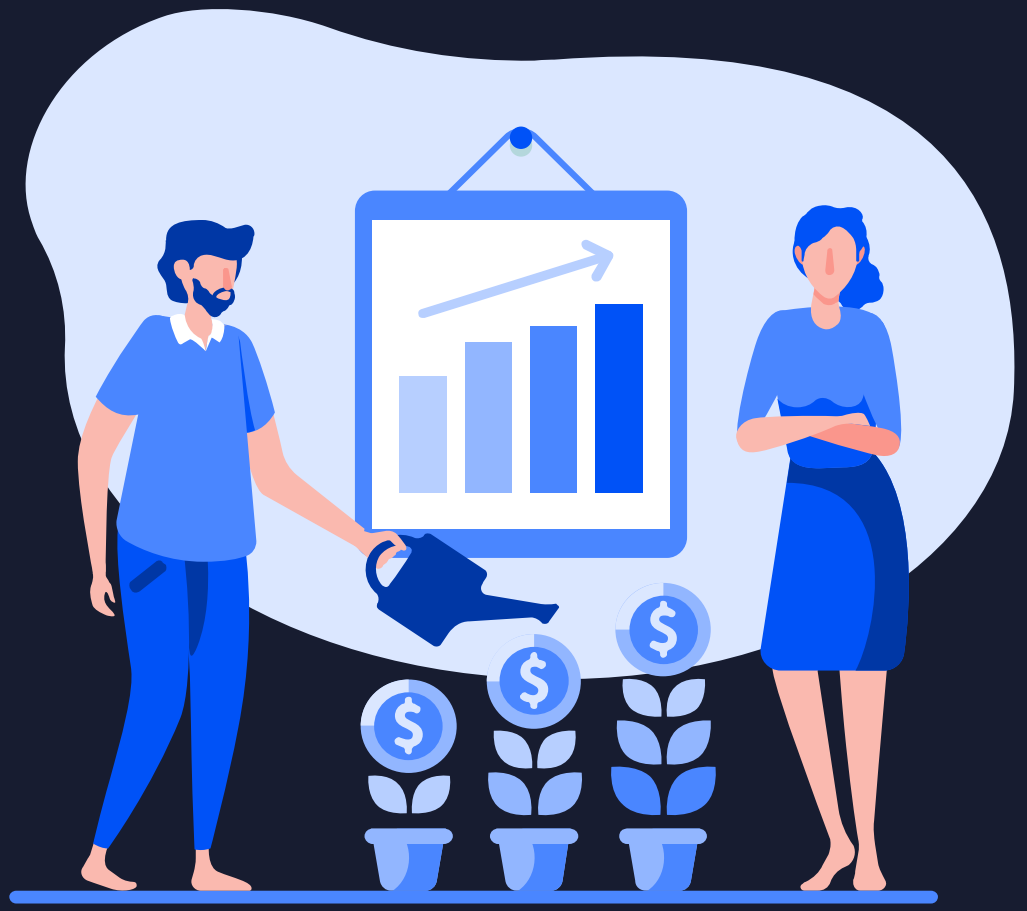
link_trace_cs.referrer	link_trace_cs.requested_article	link_trace_cs.type_traffic	link_trace_cs.occurences
The_Long_Run_(album)	Eagles_Live	link	1322
The_Long_Run_(album)	Hotel_California_(Eagles_album)	link	654
The_Long_Run_(album)	I_Can't_Tell_You_Why	link	470
The_Long_Run_(album)	Heartache_Tonight	link	327
The_Long_Run_(album)	The_Long_Run_(song)	link	319
The_Long_Run_(album)	Timothy_B._Schmit	link	319
The_Long_Run_(album)	Eagles_(band)	link	309
The_Long_Run_(album)	In_the_City_(Joe_Walsh_song)	link	297
The_Long_Run_(album)	Don_Felder	link	285
The_Long_Run_(album)	Long_Road_Out_of_Eden	link	168
The_Long_Run_(album)	Joe_Walsh	link	128

q3

RESULTS

```
Hotel_California(2222) -> Hotel_California_(Eagles_Album)(2010) -> The_Long_Run_(album)(2127) ->
Eagles_Live(1322) -> Eagles_Greatest_Hits,_Vol._2(1136) -> The_Very_Best_of_the_Eagles(996) ->
Hell_Freezes_Over(892) -> Selected_Works:_1972–1999(735) -> The_Very_Best_Of_(Eagles_album)(705) ->
Eagles_(box_set)(646)
```

what articles are relatively more popular in US, UK & Australia?



q4

PROCESS

- Sampled pageview data from Sept 1st 2020
- Used a time zone converter to get US, UK and Australian city times respectively
- Took 5-hour sample that remained the same for each chosen city for each country.

q4

PROCESS

UTC, Time Zone (UTC +0)	Tue, Sep 1, 2020	17:00
New York, NY, USA* EDT (UTC -4)	Tue, Sep 1, 2020	1:00 pm
London, United Kingdom* BST (UTC +1)	Tue, Sep 1, 2020	6:00 pm
Brisbane, Australia AEST (UTC +10)	Wed, Sep 2, 2020	3:00 am

q4

PROCESS

- For 5-hour blocks for each country, I placed them in to hive tables.

```
0: jdbc:hive2://> CREATE TABLE us_views_final AS
. . . . . . . . > SELECT ARTICLE, SUM(VIEWS) AS TOTAL
. . . . . . . . >     FROM us_views
. . . . . . . . >     WHERE DOMAIN="en" OR DOMAIN="en.m"
. . . . . . . . >     GROUP BY ARTICLE
. . . . . . . . >     SORT BY TOTAL DESC
. . . . . . . . >     LIMIT 20;
```

q4

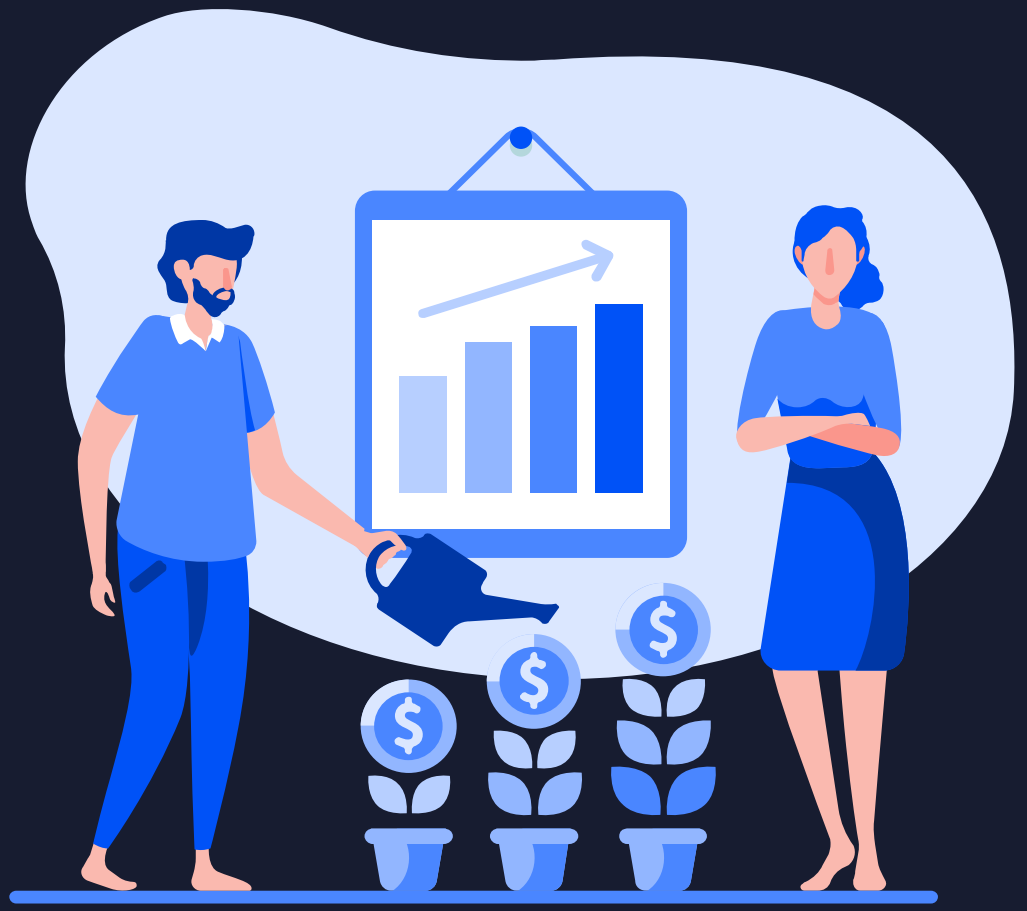
RESULTS

us_views_final.article	us_views_final.total
Main_Page	1339423
Special:Search	338980
-	148817
Chadwick_Boseman	101445
Jackie_Ormes	99931
Tenet_(film)	47633
F5_Networks	44592
Sikhism	42092
Pranab_Mukherjee	41806
Ivan_Rakitić	35584
Cobra_Kai	30952
Deaths_in_2020	27626
Robin_Williams	26161
Ron_Jeremy	26043
Robert_F._Kennedy_Jr	25092
Niecy_Nash	25040
Mammy_Two_Shoes	22709
Bible	22253
Tenerife	20786
Gabriel_dos_Santos_Magalhães	20566

uk_views_final.article	uk_views_final.total
Main_Page	1279904
Special:Search	343650
-	138535
Chadwick_Boseman	92407
Erick_Morillo	75084
Jackie_Ormes	67351
Sikhism	55978
Tenet_(film)	47948
Democritus	43731
Sheridan_Smith	43004
Cobra_Kai	39060
Vespers	31435
Bible	30933
Deaths_in_2020	30025
Robert_F._Kennedy_Jr	28211
Vespro_della_Beata_Vergine	24595
Avengers_(2020_video_game)	23621
The_Three-Body_Problem_(novel)	23232
Andy_Murray	21539
Shooting_of_Jacob_Blake	21434

aus_views_final.article	aus_views_final.total
Main_Page	1053379
Special:Search	221074
-	106617
Chadwick_Boseman	60329
Joe_Kennedy_III	48101
Tenet_(film)	44028
Ed_Markey	38991
Cobra_Kai	38401
Kim_Clijsters	38325
Erick_Morillo	29024
Jamal_Murray	24587
Bible	24504
William_Zabka	20571
Where's_Herb?	19850
Donovan_Mitchell	19657
Joseph_P._Kennedy_II	19518
Bruce_Lee	19188
Deaths_in_2020	18546
Pranab_Mukherjee	17886
Avengers_(2020_video_game)	17678

what’s the average visits a vandalized article receives before reversed?



q5

PROCESS

- For Q5, I made some simplifications due to the denormalized state of the dataset
- Simplified to worst-case of a vandalized page before it was reversed and worst-case amount of views a page could receive

q5

RESULTS

```
0: jdbc:hive2://> select * from q5_revision_final as alias;
OK
```

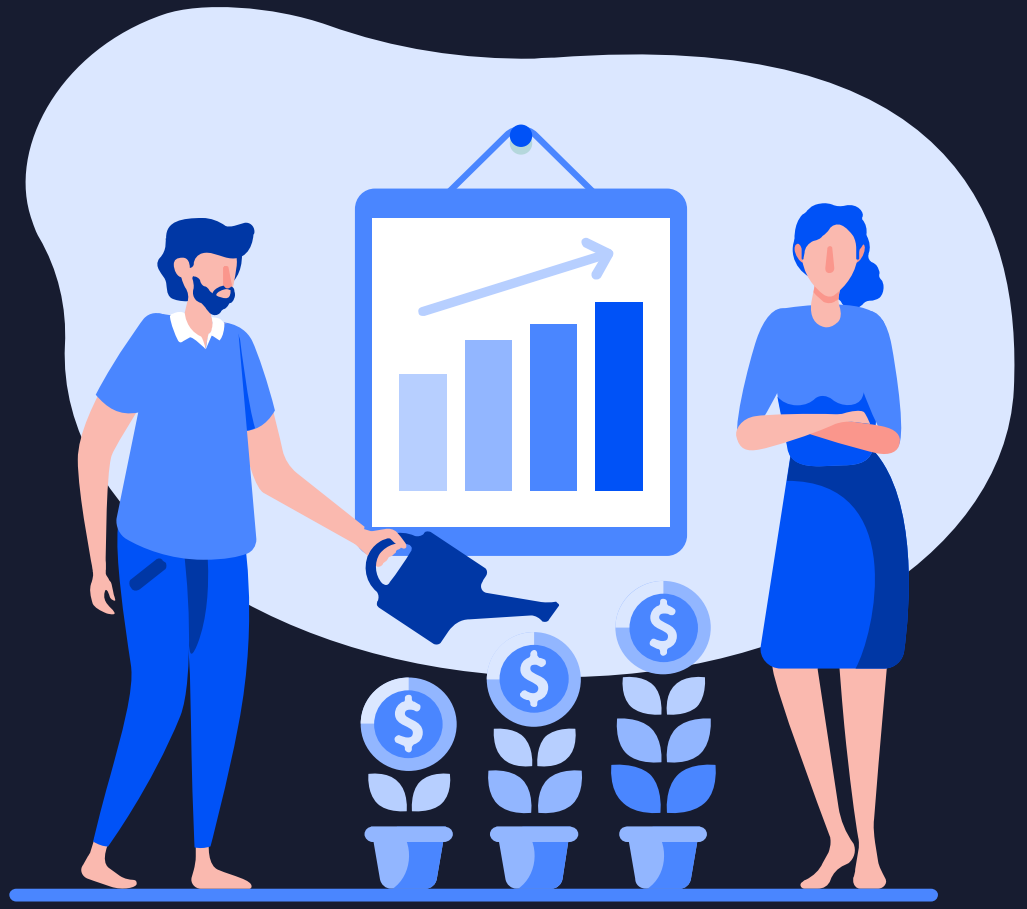
alias.db	alias.uaction	alias.type	alias.title	alias.revise_count	alias.avrg_revise_rvrse	alias.avrg_prev_rev
enwiki	revision	create	Harrah's_Reno	276	42976.05	80.12
enwiki	revision	create	Tokyo_Ghoul	2059	42534.68	7874.03
enwiki	revision	create	Steve_Carell	6487	42502.58	1256.97
enwiki	revision	create	Steve_Carell	6488	42501.87	0.72
enwiki	revision	create	Heartland_Public_Radio	52	42471.48	460800.5
enwiki	revision	create	Sasuke_Sarutobi	205	42212.73	437.98
enwiki	revision	create	Carole_Lin	71	42211.62	0.8
enwiki	revision	create	Kosovo_Security_Force	1013	41987.02	22834.68
enwiki	revision	create	HISAR_(surface_to_air_missile_system)	103	41933.98	11732.55
enwiki	revision	create	HISAR_(surface_to_air_missile_system)	104	41933.65	0.33
enwiki	revision	create	HISAR_(surface_to_air_missile_system)	105	41932.92	0.73
enwiki	revision	create	Masssly	2724	41817.28	1424.5
enwiki	revision	create	Tina_in_the_Sky_with_Diamonds	155	41766.12	35413.65
enwiki	revision	create	Smart_city	1443	41629.87	7673.27
enwiki	revision	create	Wayne,_Nebraska	189	41495.62	30850.33
enwiki	revision	create	Wayne,_Nebraska	190	41495.1	0.52
enwiki	revision	create	Wayne,_Nebraska	191	41494.5	0.6
enwiki	revision	create	WikiProject_Deletion_sorting/Theatre	880	41441.25	2558.07
enwiki	revision	create	History_of_Bremen	10	41412.25	15378.2
enwiki	revision	create	List_of_Test_cricket_records	255	41180.1	3384.88
enwiki	revision	create	Grammaticalization	341	41162.9	21689.33
enwiki	revision	create	List_of_Test_cricket_records	256	41069.17	110.93
enwiki	revision	create	List_of_English_prepositions	1902	41044.52	10415.3
enwiki	revision	create	HISAR_(surface_to_air_missile_system)	106	41008.57	924.35
enwiki	revision	create	List_of_Test_cricket_records	257	41006.3	62.87

25 rows selected (0.125 seconds)

```
0: jdbc:hive2://> select * from mrviews limit 20;
OK
```

mrviews.article	mrviews.total_views
Main_Page	165044119
Special:Search	41915305
-	17237713
Ruth_Bader_Ginsburg	7605356
Amy_Coney_Barrett	5924508
Tenet_(film)	3877047
Shooting_of_Breonna_Taylor	3850524
Dennis_Nilsen	3564441
Deaths_in_2020	3316200
Mulan_(2020_film)	3239724
The_Boys_(2019_TV_series)	3184665
Bible	3170711
Joe_Biden	2740959
Ratched_(TV_series)	2626716
Cobra_Kai	2459988
Chadwick_Boseman	2417875
S._P._Balasubrahmanyam	2387782
Microsoft_Office	2136261
XXXX	2056847
September_11_attacks	2028774

what does the mobile vs. desktop pageviews show us? possible explanations?



q6

PROCESS

- For Q6, I decided to compare the mobile and desktop pageviews
- In a company this can be important to understand the type of devices their visitors use and can help them better display information

q6

RESULTS

OK

d_code	d_article	d_views	m_code	m_article	m_views
en	Main_Page	74651691	en.m	Main_Page	90392428
en	Special:Search	25326649	en.m	Special:Search	16588656
en	-	4951915	en.m	-	12285798
en	Bible	3096095	en.m	Bible	74616
en	Deaths_in_2020	1751575	en.m	Deaths_in_2020	1564625
en	Ruth_Bader_Ginsburg	1691377	en.m	Ruth_Bader_Ginsburg	5913979
en	Amy_Coney_Barrett	1658322	en.m	Amy_Coney_Barrett	4266186
en	Mulan_(2020_film)	1581302	en.m	Mulan_(2020_film)	1658422
en	Tenet_(film)	1568600	en.m	Tenet_(film)	2308447
en	F5_Networks	1362925	en.m	F5_Networks	125030
en	The_Boys_(2019_TV_series)	1236556	en.m	The_Boys_(2019_TV_series)	1948109
en	Portal:Current_events	924128	en.m	Portal:Current_events	310674
en	Shooting_of_Breonna_Taylor	803933	en.m	Shooting_of_Breonna_Taylor	3046591
en	Periodic_table	785868	en.m	Periodic_table	590826
en	Raised_by_Wolves_(American_TV_series)	774277	en.m	Raised_by_Wolves_(American_TV_series)	994614
en	Microsoft_Office	756674	en.m	Microsoft_Office	1379587
en	YouTube	751802	en.m	YouTube	537775
en	Joe_Biden	721011	en.m	Joe_Biden	2019948
en	COVID-19_pandemic	691609	en.m	COVID-19_pandemic	684730
en	COVID-19_pandemic_by_country_and_territory	689999	en.m	COVID-19_pandemic_by_country_and_territory	517881


questions?

q1

QUERIES

```
0: jdbc:hive2://> SELECT ARTICLE_NAME, SUM(PAGE_VIEWS) AS TOTAL
. . . . . . . . >     FROM PAGEVIEWS
. . . . . . . . >     WHERE DOMAIN_CODE="en" OR DOMAIN_CODE="en.m"
. . . . . . . . >     GROUP BY ARTICLE_NAME
. . . . . . . . >     SORT BY TOTAL DESC
. . . . . . . . >     LIMIT 10;
```

```
0: jdbc:hive2://> SELECT DOMAIN_CODE, ARTICLE_NAME, SUM(PAGE_VIEWS) AS TOTAL
. . . . . . . . >     FROM PAGEVIEWS
. . . . . . . . >     WHERE DOMAIN_CODE="en" OR DOMAIN_CODE="en.m"
. . . . . . . . >     GROUP BY DOMAIN_CODE, ARTICLE_NAME
. . . . . . . . >     SORT BY TOTAL DESC
. . . . . . . . >     LIMIT 10;
```

q2

QUERIES

```
0: jdbc:hive2://> CREATE TABLE MRVIEWS AS
. . . . . . . . > SELECT ARTICLE, SUM(PAGE_VIEWS) AS TOTAL_VIEWS
. . . . . . . . >     FROM PRE_MRVIEWS
. . . . . . . . >     WHERE DOMAIN_CODE="en" OR DOMAIN_CODE="en.m"
. . . . . . . . >     GROUP BY ARTICLE
. . . . . . . . >     SORT BY TOTAL_VIEWS DESC
. . . . . . . . >     LIMIT 25;
```

```
0: jdbc:hive2://> CREATE TABLE Q2_MERGED_MR AS
. . . . . . . . >     SELECT MRVIEWS.ARTICLE, MRVIEWS.TOTAL_VIEWS, MRCLICKS.INTERNAL_LINK
. . . . . . . . >     FROM MRCLICKS
. . . . . . . . >     JOIN MRVIEWS ON (MRVIEWS.ARTICLE = MRCLICKS.ARTICLE)
. . . . . . . . >     LIMIT 25;
```

```
0: jdbc:hive2://> CREATE TABLE Q2_RESULTS AS
. . . . . . . . >     SELECT *, CAST (INTERNAL_LINK / TOTAL_VIEWS * 100 AS DECIMAL(5, 3)) AS PERCENT_INTERNAL
. . . . . . . . >     FROM Q2_MERGED_MR
. . . . . . . . >     SORT BY PERCENT_INTERNAL DESC
. . . . . . . . >     LIMIT 10;
```

q3

QUERIES

```
0: jdbc:hive2://> select * from link_trace_cs
. . . . . . . . >     where type_traffic="link" and  referrer like "Hotel_California" and not (referrer = "other-internal" or referrer= "other-search" or
. . . . . . . . >     referrer="other-external"or referrer="other-empty"or referrer ="other-other")
. . . . . . . . >     sort by occurences desc
. . . . . . . . >     limit 100;
```

```
. . . . . . . . >     where type_traffic="link" and referrer like "Hotel_California_(Eagles_album)" and not (referrer = "other-internal" or referrer
= "other-search" or
. . . . . . . . >     referrer="other-external"or referrer="other-empty"or referrer ="other-other")
. . . . . . . . >     sort by occurences desc
. . . . . . . . >     limit 100;
```

q5

QUERIES

```
create table q5_revision_final as
select WIKI_DB AS DB, EVENT_ENTITY AS UAction, EVENT_TYPE as type,
        page_title AS TITLE, page_revision_count AS Revise_Count,
        Round(AVG(Distinct revision_seconds_to_identity_revert/60),2) AS Avrg_Revise_Rvrse,
        Round(AVG(Distinct page_seconds_since_previous_revision/60),2) AS Avrg_Prev_Rev FROM revisions
        where revision_seconds_to_identity_revert > 0
        and page_seconds_since_previous_revision > 0
        GROUP By Wiki_db, EVENT_ENTITY, event_type, page_title, page_revision_count,
        revision_seconds_to_identity_revert, page_seconds_since_previous_revision
        sort by Avrg_Revise_Rvrse desc
        limit 25;
```