# Project 1

Sean Tidd

# Project Goals

- Process multiple large datasets from Wikipedia and get discussable results
- Use a combination of Hadoop MapReduce and Hive's query ability to get outputs for project questions
- Design the program/structure efficiently

# Project Design

The project design consisted of 3 steps:

1. MapReduce Raw Data
    a. Scala jar file in Hadoop's MapReduce
    b. Heavy computation on MapReduce to create/join/query less on Hive tables
2. Create Hive Tables From MapReduce Output
    a. MapReduce output is converted to a Hive table
    b. External tables to prevent repetition of MapReduce data upload
3. Query Hive Tables
    a. Queries designed to be simple by assuming the previous steps simplified the process

# MapReduce

1. **MapReduce Jar Q1**
    a. **Input File Type:** pageviews files
    b. **Outputs:** file of key-value pairs of the wiki title and the total views
    c. **Other Features:** filters out everything that is not in the domain of “en” or “en.m
2. **MapReduce Jar Q2**
    a. **Input File Type:** clickstream files
    b. **Outputs:** file of key-value pairs of the wiki title and the total users that followed a link
    c. **Other Features:** filters out everything that is not of “link” type

# MapReduce (continued)

3. **MapReduce Jar Q3**
    a. **Input File Type:** clickstream files
    b. **Outputs:** file of key-value pairs of the wiki title and its maximum value pair (link title and the total views)
    c. **Other Features:** filters out everything that is not “link” type

# Difficulties

- The Data Format
  - A lot of the data is not user friendly (especially true for the wiki edits)
  - Many useful fields like country origin were missing (even with domain, en is for multiple countries for example)
- The Data Volume
  - The necessary data for the pageviews for the FULL month of September was massive and comprised of 720 zip files
- A Cluster with a Single Machine
  - All 720 zip files were stored and evaluated through MapReduce on a single machine, eating up storage and taking hours as MapReduce processed a result

# Project Results

**Q1:** Which English wikipedia article got the most traffic on October 20?

| wiki_oct_total_views.wiki_title | wiki_oct_total_views.total_views |
|---|---|
| Main_Page | 11922016 |
| Special:Search | 2953662 |
| - | 1089428 |
| Jeffrey_Toobin | 642918 |
| C._Rajagopalachari | 421116 |
| The_Haunting_of_Bly_Manor | 370278 |
| Robert_Redford | 357558 |
| Jeff_Bridges | 318326 |
| Bible | 302968 |
| Chicago_Seven | 299932 |

10 rows selected (29.074 seconds)

**Assumptions:** Used both en and en.m domains.

# Project Results (continued)

**Q2:** What English wikipedia article has the largest fraction of its readers follow an internal link to another wikipedia article?

| view_link_fraction.wiki_title | view_link_fraction.total_views | view_link_fraction.link_total | view_link_fraction.percentage |
|---|---|---|---|
| List_of_neo-Nazi_bands | 7256 | 7255 | 99.98621830209483 |
| HM_Prison_Manchester | 6938 | 6937 | 99.98558662438744 |
| Kygo_discography | 6123 | 6122 | 99.98366813653438 |
| Life_(1999_film) | 13657 | 13654 | 99.97803324302555 |
| Ka_(pharaoh) | 4422 | 4421 | 99.97738579828132 |
| Shootout | 3652 | 3651 | 99.97261774370209 |
| ABC | 3600 | 3599 | 99.97222222222221 |
| 2008_Tour_de_France | 10057 | 10054 | 99.9701700308243 |
| List_of_restaurant_terminology | 6584 | 6582 | 99.96962332928311 |
| My_Chemical_Romance_discography | 6515 | 6513 | 99.96930161166539 |

**Assumptions:** Ignored any values past 100% and ignored any values at exactly 100% (ignore small values such as 1 total view and 1 total links, etc.). Used both en and en.m domains.

# Project Results (continued)

**Q3:** What series of wikipedia articles, starting with Hotel California, keeps the largest fraction of its readers clicking on internal links? This is similar to (2), but you should continue the analysis past the first article.

| series_path.wiki_title | series_path.link | series_path.fraction |
|---|---|---|
| Hotel_California | Hotel_California_(Eagles_album) | 0.04584468 |
| Hotel_California_(Eagles_album) | The_Long_Run_(album) | 0.10107394 |
| The_Long_Run_(album) | Eagles_Live | 0.13885096 |
| Eagles_Live | Eagles_Greatest_Hits,_Vol._2 | 0.35148516 |
| Eagles_Greatest_Hits,_Vol._2 | The_Very_Best_of_the_Eagles | 0.45149592 |

**Assumptions:** Ignored any values past 100% and ignored any values at exactly 100% (ignore small values such as 1 total view and 1 total links, etc.). Used both en and en.m domains. The user looks up the next link as a where clause to add to the Series Path(Hadoop is not acyclic to perform loops).

# Project Results (continued)

**Q4:** Find an example of an English wikipedia article that is relatively more popular in the UK. Find the same for the US and Australia.

| final_edit_popularity.page_title | final_edit_popularity.region | final_edit_popularity.edit_popularity |
|---|---|---|
| List_of_Joe_Biden_2020_presidential_campaign_endorsements | UK | 521 |
| Gioguch/sandbox | UK | 459 |
| American_Revolutionary_War | UK | 418 |

| final_edit_popularity.page_title | final_edit_popularity.region | final_edit_popularity.edit_popularity |
|---|---|---|
| Administrator_intervention_against_vandalism | US | 3274 |
| Teahouse | US | 1548 |
| Requests_for_page_protection | US | 1515 |

| final_edit_popularity.page_title | final_edit_popularity.region | final_edit_popularity.edit_popularity |
|---|---|---|
| Sandbox | AU | 1336 |
| WQUlrich/sandbox | AU | 754 |
| Swaminarayan_Sampradaya | AU | 488 |

**Assumptions:** Popularity is with respect to highest frequency of edits. Edits made within the internet rush hour(7pm-11pm UTC) during a country's specific time zone is assumed to be originated from that country.

# Project Results (continued)

**Q5:** Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.

| wiki_vandals.page_title_historical | wiki_vandals.revision_seconds_to_identity_revert | wiki_vandals.revision_days_to_identity_revert | wiki_vandals.revert_month_fraction_time |
|---|---|---|---|
| Uncharted_2:_Among_Thieves | 29 | 3.356481481481481E-4 | 1.1188271604938271E-5 |
| List_of_governors_of_Louisiana | 369778 | 4.279837962962963 | 0.14266126543209878 |
| Uncharted_2:_Among_Thieves | 6 | 6.944444444444444E-5 | 2.3148148148148148E-6 |
| GlowSmokey/sandbox | 363 | 0.004201388888888889 | 1.400462962962963E-4 |
| Cyberbot_I/Status | 86403 | 1.0000347222222221 | 0.033334490740740734 |
| Filters_in_topology | 31 | 3.587962962962963E-4 | 1.1959876543209876E-5 |
| Cyberbot_II/Status | 86402 | 1.000023148148148 | 0.0333341049382716 |
| Ratched_(TV_series) | 1931 | 0.022349537037037036 | 7.449845679012346E-4 |
| Telepace | 915 | 0.010590277777777778 | 3.530092592592593E-4 |
| Unknown_Hinson | 32 | 3.7037037037037035E-4 | 1.2345679012345678E-5 |

**Assumptions:** Used only “en” domain. Assumed that any article that had a revision reverted was do to vandalism. Rough estimation on number of views is in the following expression: *days_until_revert* = (*seconds_until_revert*)/ 86400

*revert_month_fraction_time* = (*days_until_revert*)/30

*vandalized_page_viewers* = (*revert_month_fraction_time*)*(*total_september_page_views*)

| avg_vandal_page_viewers |
|---|
| 399.7598655534004 |

# Project Results (continued)

**Q6:** What English Wikipedia article is the closest to having half of it's viewers clicking on internal links to other articles without being exactly 50%?

| half_links.wiki_title | half_links.percentage | half_links.closeness_to_half |
|---|---|---|
| Keladi_Kanmani | 49.998258749782345 | 0.0017412502176554767 |
| Leaving_Neverland | 49.997793760755414 | 0.0022062392445860723 |
| 2020_Sabah_state_election | 50.00250551212668 | 0.002505512126681708 |
| John_Barrymore | 50.0027481083854 | 0.0027481083853970745 |
| 1971_Bangladesh_genocide | 50.00304154753938 | 0.00304154753938235 |
| Brian_Christopher | 49.996755985207294 | 0.0032440147927061957 |
| Xiao_Zhan | 49.99582707394425 | 0.004172926055751702 |
| Mike_Myers | 49.99495204442201 | 0.00504795557799298 |
| Jack_McVitie | 50.00535618639529 | 0.005356186395289342 |
| Thoroughbreds_(2017_film) | 50.00555308751666 | 0.005553087516659616 |

**Assumptions:** Ignored any values past 100% and ignored any values at exactly 100% (ignore small values such as 1 total view and 1 total links, etc.). Used both en and en.m domains.

# Thank You