Report

The Problem

Providing Wikipedia page view statistics in an accessible and size minimized format:

accessible

- searchable
- structured in time series

minimized

- removing data redundancies
- with minimal amount of bytes/characters
- compress?

The task at hand is a typical big data problem. The amount of data points is so large that key decisions will scale enormously having huge effects on data structure, network connection at hand, hardware, hardware costs, time costs and time constraints. Furthermore, many traditional approaches - like simply using more hardware - might simply be not feasible. Second, developing an execution plan that brings together goals (accessible data, in 'portable' size), constraints (cost, time) and options is vital. Therefore, a planning phase was needed.

Data

Granularity

There are two types of granularity available, daily aggregates and hourly aggregates:

- daily: 2011-11 - present (https://dumps.wikimedia.org/other/pagecounts-ez/merged/)
- hourly: 2007 - 2016 (https://dumps.wikimedia.org/other/pagecounts-raw/)

Further data

- page titles: 2018 - present (https://dumps.wikimedia.org/other/pagetitles/)

Size

Based on the daily/hourly aggregates there follow some extrapolations. In regard to network traffic (downloads), uncompromising and filtering hourly data is 24 times harder to come by than for the daily aggregates.

- more than 200 GB per year in compressed form –> (4 * 200 + 4) * (200 * 24) = **20 TB in total**
- one month needs 3 hours to download –> (3 * 12 * 4) + (3 * 12 * 24 * 4) = 6 days + 144 days hours in total
- one day needs 3 Minutes to uncompressed and filter –> ((3 * 365 * 4) + (3 * 365 * 4 * 24)) / (60 * 24) = **76 days**
- about 5 GB / 120 GB per day uncompressed –> (365 * 5 * 4) + (365 * 5 * 24 * 4) + = **182.5 TB**
- size in database 10 GB per day –> 365 * 10 * 8 = **29 TB** (naive size!)

Most of the size stems from the request titles which are repeated for each day (each hour for the more granular data format) e.g. 630 MB for Germany and from the fact that the titles stored are requests send by users (e.g. 630 MB for German Wikipedia at 2015-01-02) instead of available pages (63 MB for German Wikipedia at 2018-05-01).

In regard to storage size hourly aggregates do not take up more space once they have been put into a proper format without unnecessary redundancies.

Quality

- data is byte wise storage of requests
- not per page
- not cleaned
- different encodings (UTF8, latin1, . . .)
- URL-encoding
- non sensible requests

Actions taken so far

- download of all 2015 data > 200 GB
- putting whole day for all languages into database
- putting several days for German and English Wikipedia into database
- trying out different strategies and technologies to unpack and filter data
  - R
  - GNU-tools
- trying out, planing and calculating different hardware approaches
  - AWS
  - own server
  - own local desktop
  - dedicated PC
- planning a data structure to match goals
  - size
  - accessibility
- developing data management process and software prototype to
  - extract
  - filter
  - upload

Problems ahead and bottlenecks

The real challenge is to process hourly data due to download and uncompress / filtering times needed. Those will have to be parallelized across computers with good internet connections.

For processing daily files no further hardware is needed - only for storing and delivering end results.

For processing hourly data a lot more hardware and network bandwidth is needed: 4 to 6 times the test system. Furthermore, processing hourly data will need much more attention distributing and monitoring the execution. On the upside it seems that a lot of work in regard to processing hourly data has already been done by GESIS (https://github.com/gesiscss/wiki-download-parse-page-views) - still those procedures have to be adapted to fit into a general framework.

hardware

- SSD for storage of results: ~ 100 €
- processing server: 6 times 50 € ~ 300 €
- (unforeseen hardware needs 0 - 600 €)

total: 400 - 1000 €

manpower

- building database structure and do database administration: **16 h**
- processing daily data: **16 h**
- processing hourly data: **25 h**
- IT administration, research and orchestration **10 h**

total: 67 h (2680)

optional

- building interface for data access with wikipediatrend compatibility: 10h
- data cleanup: 25 h plus (has to be evaluated separately)