README.md

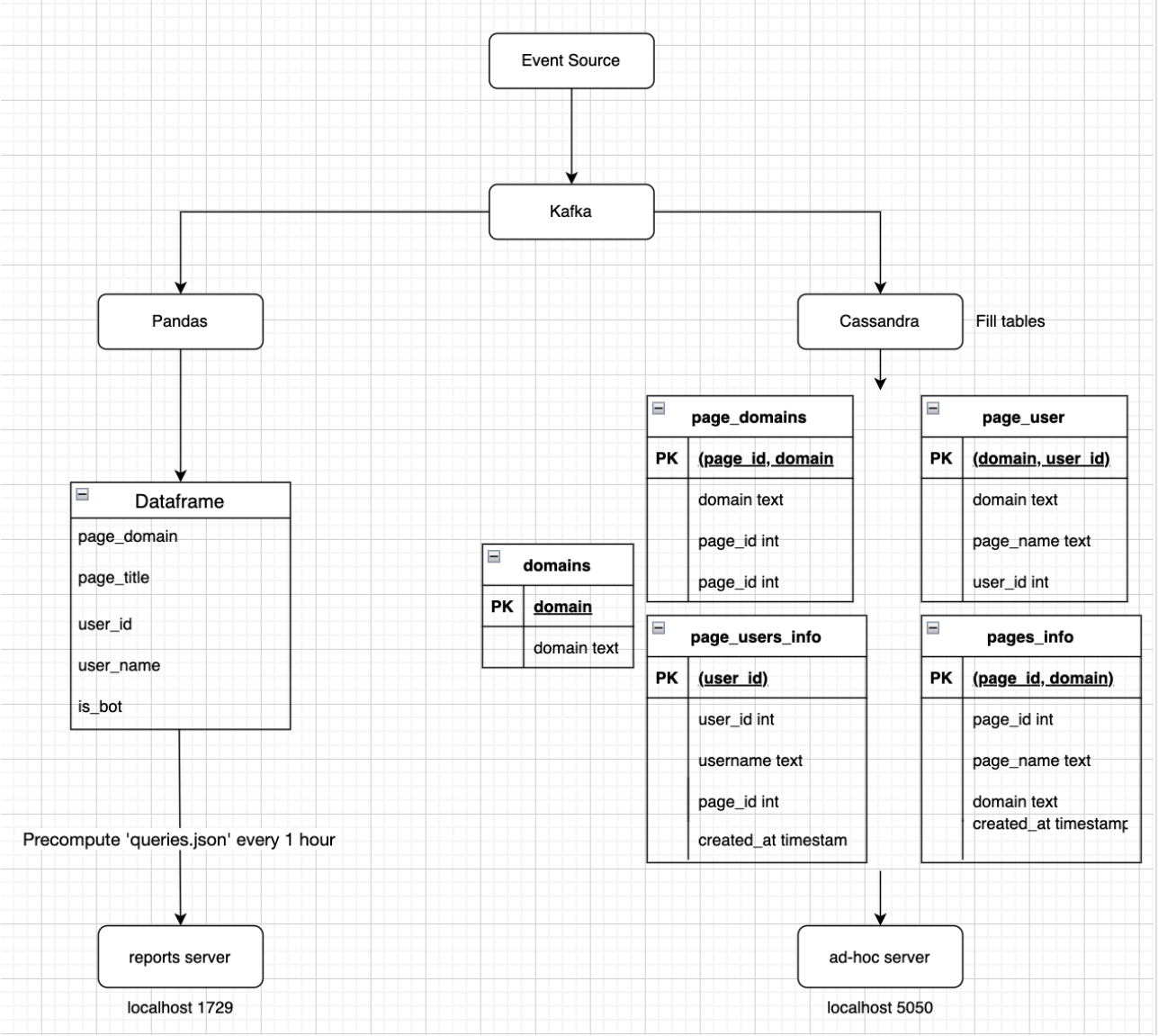
Course project - Big Data Processing

Team

- Yarema Mishchenko
- Dmytro Lutchyn

System design

Diagram



Kafka - producer & consumer

TODO: (yarema) write about how these work in our system

Cassandra - DB for storage

TODO: (dmytro) write about tables in Cassandra

Precomputed reports server - Category A

TODO: (yarema) write about why pandas and how this server gets data

Ad hoc server - Category B

TODO: (dmytro) write about how this server gets data

Results

All of our results are stored in the `project-results` directory. These results are generated using the `client_demo.py` script. As per the duration requirement, the system was running for ~8h 27m before making the requests.

Category A

For category A, the `.json` file with query results is saved to `project-results/queries.json`.

The file is extremely large because of the 3rd query: the top author created over 20000 pages.

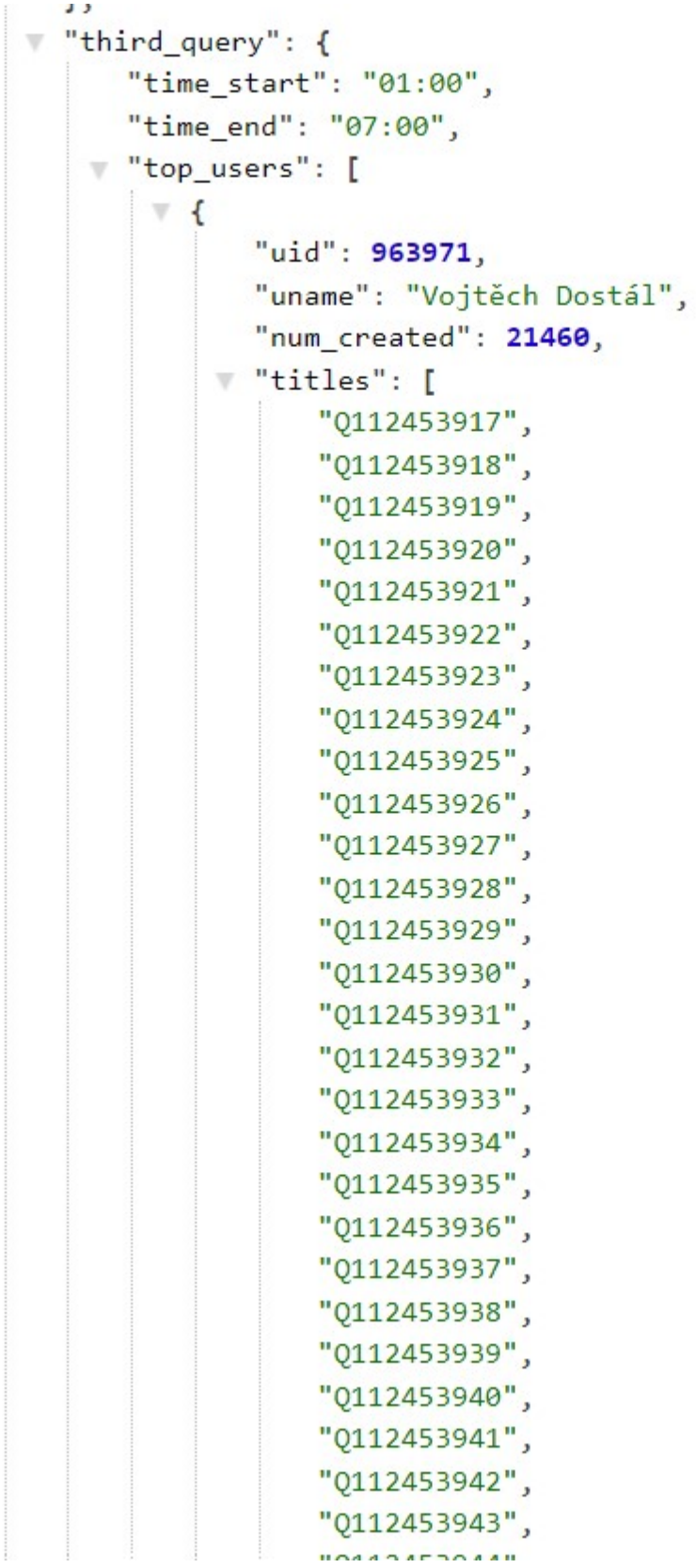
First query:

```
"first_query": [
  {
    "time_start": "01:00",
    "time_end": "02:00",
    "statistics": {
      "ar.wikipedia.org": 197,
      "arz.wikipedia.org": 99,
      "ban.wikipedia.org": 1,
      "bg.wikipedia.org": 2,
      "bjn.wikipedia.org": 1,
      "ca.wikipedia.org": 1,
      "ceb.wikipedia.org": 1,
      "commons.wikimedia.org": 1010,
      "cy.wikisource.org": 1,
      "da.wikipedia.org": 14,
      "de.wikipedia.org": 4,
      "en.wikibooks.org": 1,
      "en.wikinews.org": 4,
      "en.wikipedia.org": 179,
      "en.wikiquote.org": 1,
      "en.wikisource.org": 121,
      "en.wiktionary.org": 34,
      "es.wikipedia.org": 34,
      "es.wikiquote.org": 1,
      "es.wikisource.org": 4,
      "es.wiktionary.org": 1,
      "eu.wikipedia.org": 1,
      "fa.wikipedia.org": 34,
      "fi.wikipedia.org": 2,
      "fr.wikipedia.org": 14,
      "fr.wikisource.org": 18,
      "gu.wikipedia.org": 1,
      "hi.wikipedia.org": 1,
      "hr.wikipedia.org": 1,
      "ht.wikipedia.org": 2,
      "hu.wikipedia.org": 2,
      "id.wikipedia.org": 4,
      "incubator.wikimedia.org": 7,
      "is.wikipedia.org": 2,
```


Second query:

```
"second_query": {
    "time_start": "01:00",
    "time_end": "07:00",
    "statistics": [
        {
            "domain": "ar.wikipedia.org",
            "created_by_bots": 392
        },
        {
            "domain": "as.wikipedia.org",
            "created_by_bots": 17
        },
        {
            "domain": "bn.wikipedia.org",
            "created_by_bots": 2
        },
        {
            "domain": "ca.wikipedia.org",
            "created_by_bots": 1
        },
        {
            "domain": "commons.wikimedia.org",
            "created_by_bots": 3407
        },
        {
            "domain": "de.wikipedia.org",
            "created_by_bots": 10
        },
        {
            "domain": "en.wikipedia.org",
            "created_by_bots": 65
        },
        {
            "domain": "es.wikipedia.org",
            "created_by_bots": 2
        },
        {
            "domain": "eu.wikipedia.org",
            "created_by_bots": 7
```

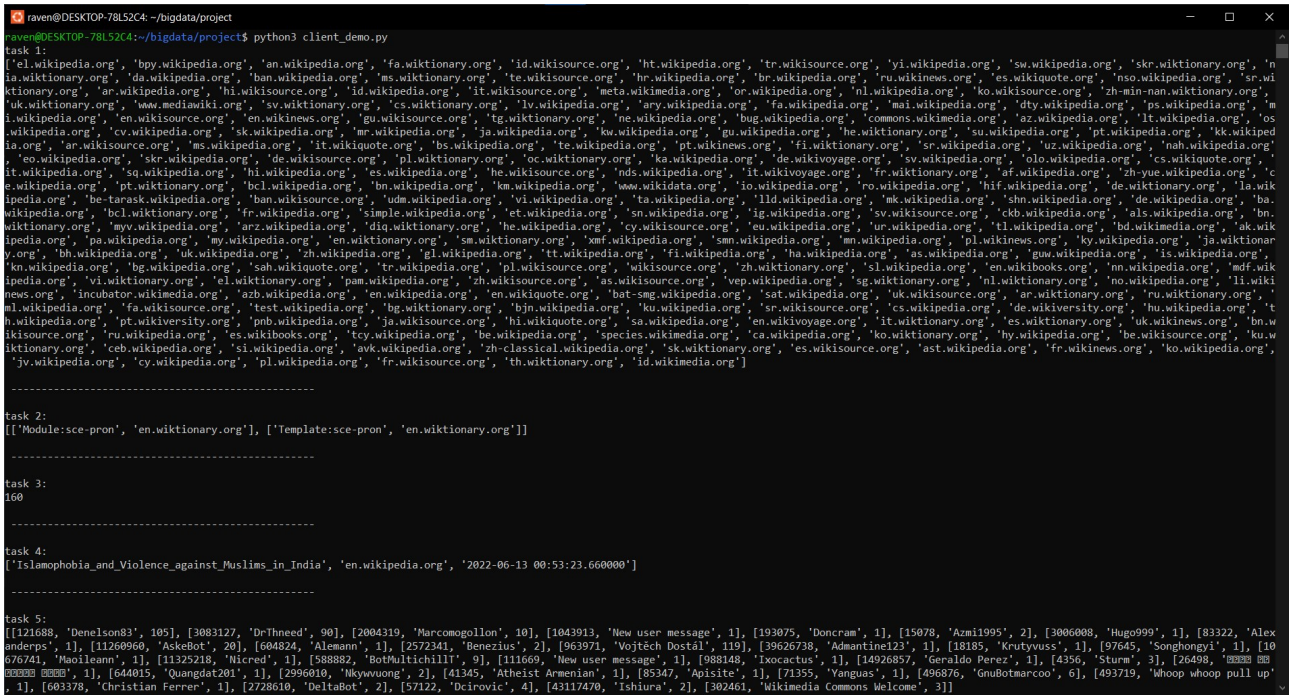

Third query:



Category B

As for category B, the output is displayed directly when running `client_demo.py` .

The arguments were chosen to fit well in a screen.


raven@DESKTOP-78L52C4: ~/bigdata/project
raven@DESKTOP-78L52C4:~/bigdata/project$ python3 client_demo.py
task 1:
['el.wikipedia.org', 'bpy.wikipedia.org', 'an.wikipedia.org', 'fa.wiktionary.org', 'id.wikisource.org', 'ht.wikipedia.org', 'tr.wikisource.org', 'yi.wikipedia.org', 'sw.wikipedia.org', 'skr.wiktionary.org', 'nia.wiktionary.org', 'da.wikipedia.org', 'ban.wikipedia.org', 'ms.wiktionary.org', 'te.wikisource.org', 'hr.wikipedia.org', 'br.wikipedia.org', 'ru.wikinews.org', 'es.wikiquote.org', 'nso.wikipedia.org', 'sr.wiktionary.org', 'ar.wikipedia.org', 'hi.wikisource.org', 'id.wikipedia.org', 'it.wikisource.org', 'meta.wikimedia.org', 'or.wikipedia.org', 'nl.wikipedia.org', 'ko.wikisource.org', 'zh-min-nan.wiktionary.org', 'uk.wiktionary.org', 'www.mediawiki.org', 'sv.wiktionary.org', 'cs.wiktionary.org', 'lv.wikipedia.org', 'ary.wikipedia.org', 'fa.wikipedia.org', 'mai.wikipedia.org', 'dty.wikipedia.org', 'ps.wikipedia.org', 'mi.wikipedia.org', 'en.wikisource.org', 'en.wikinews.org', 'gu.wikisource.org', 'tg.wiktionary.org', 'ne.wikipedia.org', 'bug.wikipedia.org', 'commons.wikimedia.org', 'az.wikipedia.org', 'lt.wikipedia.org', 'os.wikipedia.org', 'cv.wikipedia.org', 'sk.wikipedia.org', 'mr.wikipedia.org', 'ja.wikipedia.org', 'kw.wikipedia.org', 'gu.wikipedia.org', 'he.wiktionary.org', 'su.wikipedia.org', 'pt.wikipedia.org', 'kk.wikipedia.org', 'ar.wikisource.org', 'ms.wikipedia.org', 'it.wikiquote.org', 'bs.wikipedia.org', 'te.wikipedia.org', 'pt.wikinews.org', 'fi.wiktionary.org', 'sr.wikipedia.org', 'uz.wikipedia.org', 'nah.wikipedia.org', 'eo.wikipedia.org', 'skr.wikipedia.org', 'de.wikisource.org', 'pl.wiktionary.org', 'oc.wiktionary.org', 'ka.wikipedia.org', 'de.wikivoyage.org', 'sv.wikipedia.org', 'olo.wikipedia.org', 'cs.wikiquote.org', 'it.wikipedia.org', 'sq.wikipedia.org', 'hi.wikipedia.org', 'es.wikipedia.org', 'he.wikisource.org', 'nds.wikipedia.org', 'it.wikivoyage.org', 'fr.wiktionary.org', 'af.wikipedia.org', 'zh-yue.wikipedia.org', 'ce.wikipedia.org', 'pt.wiktionary.org', 'bcl.wikipedia.org', 'bn.wikipedia.org', 'km.wikipedia.org', 'www.wikidata.org', 'io.wikipedia.org', 'ro.wikipedia.org', 'hif.wikipedia.org', 'de.wiktionary.org', 'la.wikipedia.org', 'be-tarask.wikipedia.org', 'ban.wikisource.org', 'udm.wikipedia.org', 'vi.wikipedia.org', 'ta.wikipedia.org', 'lld.wikipedia.org', 'mk.wikipedia.org', 'shn.wikipedia.org', 'de.wikipedia.org', 'ba.wikipedia.org', 'bcl.wiktionary.org', 'fr.wikipedia.org', 'simple.wikipedia.org', 'et.wikipedia.org', 'sn.wikipedia.org', 'ig.wikipedia.org', 'sv.wikisource.org', 'ckb.wikipedia.org', 'als.wikipedia.org', 'bn.wiktionary.org', 'myv.wikipedia.org', 'arz.wikipedia.org', 'diq.wiktionary.org', 'he.wikipedia.org', 'cy.wikisource.org', 'eu.wikipedia.org', 'ur.wikipedia.org', 'tl.wikipedia.org', 'bd.wikimedia.org', 'ak.wikipedia.org', 'pa.wikipedia.org', 'my.wikipedia.org', 'en.wiktionary.org', 'sm.wiktionary.org', 'xmf.wikipedia.org', 'smn.wikipedia.org', 'mn.wikipedia.org', 'pl.wikinews.org', 'ky.wikipedia.org', 'ja.wiktionary.org', 'bh.wikipedia.org', 'uk.wikipedia.org', 'zh.wikipedia.org', 'gl.wikipedia.org', 'tt.wikipedia.org', 'fi.wikipedia.org', 'ha.wikipedia.org', 'as.wikipedia.org', 'guw.wikipedia.org', 'is.wikipedia.org', 'kn.wikipedia.org', 'bg.wikipedia.org', 'sah.wikiquote.org', 'tr.wikipedia.org', 'pl.wikisource.org', 'wikisource.org', 'zh.wiktionary.org', 'sl.wikipedia.org', 'en.wikibooks.org', 'nn.wikipedia.org', 'mdf.wikipedia.org', 'vi.wiktionary.org', 'el.wiktionary.org', 'pam.wikipedia.org', 'zh.wikisource.org', 'as.wikisource.org', 'vep.wikipedia.org', 'sg.wiktionary.org', 'nl.wiktionary.org', 'no.wikipedia.org', 'li.wikinews.org', 'incubator.wikimedia.org', 'azb.wikipedia.org', 'en.wikipedia.org', 'en.wikiquote.org', 'bat-smg.wikipedia.org', 'sat.wikipedia.org', 'uk.wikisource.org', 'ar.wiktionary.org', 'ru.wiktionary.org', 'ml.wikipedia.org', 'fa.wikisource.org', 'test.wikipedia.org', 'bg.wiktionary.org', 'bjn.wikipedia.org', 'ku.wikipedia.org', 'sr.wikisource.org', 'cs.wikipedia.org', 'de.wikiversity.org', 'hu.wikipedia.org', 'th.wikipedia.org', 'pt.wikiversity.org', 'pnb.wikipedia.org', 'ja.wikisource.org', 'hi.wikiquote.org', 'sa.wikipedia.org', 'en.wikivoyage.org', 'it.wiktionary.org', 'es.wiktionary.org', 'uk.wikinews.org', 'bn.wikisource.org', 'ru.wikipedia.org', 'es.wikibooks.org', 'tcy.wikipedia.org', 'be.wikipedia.org', 'species.wikimedia.org', 'ca.wikipedia.org', 'ko.wiktionary.org', 'hy.wikipedia.org', 'be.wikisource.org', 'ku.wiktionary.org', 'ceb.wikipedia.org', 'si.wikipedia.org', 'avk.wikipedia.org', 'zh-classical.wikipedia.org', 'sk.wiktionary.org', 'es.wikisource.org', 'ast.wikipedia.org', 'fr.wikinews.org', 'ko.wikipedia.org', 'jv.wikipedia.org', 'cy.wikipedia.org', 'pl.wikipedia.org', 'fr.wikisource.org', 'th.wiktionary.org', 'id.wikimedia.org']
task 2:
[['Module:sce-pron', 'en.wiktionary.org'], ['Template:sce-pron', 'en.wiktionary.org']]
task 3:
160
task 4:
['Islamophobia_and_Violence_against_Muslims_in_India', 'en.wikipedia.org', '2022-06-13 00:53:23.660000']
task 5:
[[121688, 'Denelson83', 105], [3083127, 'DrThneed', 90], [2004319, 'Marcomogollon', 10], [1043913, 'New user message', 1], [193075, 'Doncram', 1], [15078, 'Azmi1995', 2], [3006008, 'Hugo999', 1], [83322, 'Alexanderps', 1], [11260960, 'AskeBot', 20], [604824, 'Alemann', 1], [2572341, 'Benezius', 2], [963971, 'Vojtěch Dostál', 119], [39626738, 'Admantine123', 1], [18185, 'Krutyvuss', 1], [97645, 'Songhongyi', 1], [10676741, 'Maoileann', 1], [11325218, 'Nicred', 1], [588882, 'BotMultichillT', 9], [111669, 'New user message', 1], [988148, 'Ixocactus', 1], [14926857, 'Geraldo Perez', 1], [4356, 'Sturm', 3], [26498, '[illegible]', 1], [644015, 'Quangdat201', 1], [2996010, 'Nkywvuong', 2], [41345, 'Atheist Armenian', 1], [85347, 'Apisite', 1], [71355, 'Yanguas', 1], [496876, 'GnuBotmarcoo', 6], [493719, 'Whoop whoop pull up', 1], [603378, 'Christian Ferrer', 1], [2728610, 'DeltaBot', 2], [57122, 'Dcirovic', 4], [43117470, 'Ishiura', 2], [302461, 'Wikimedia Commons Welcome', 3]]