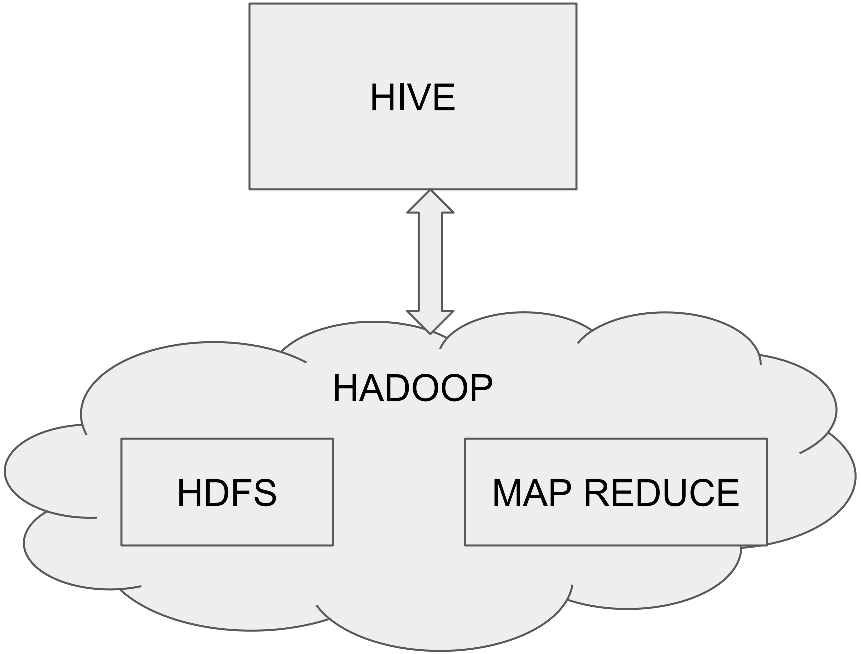
Introduction to Hive

Madhumitha Viswanathan

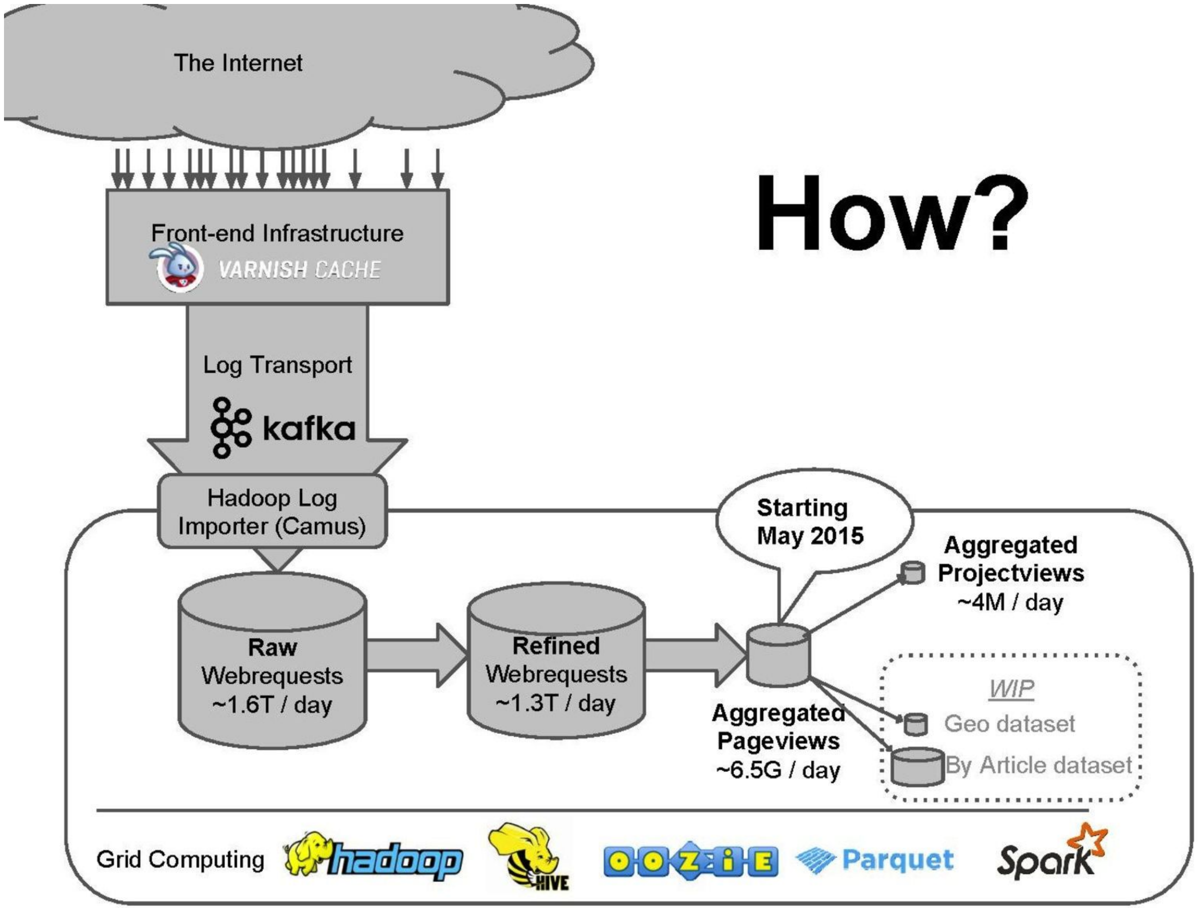
@madhuvishy

#wikimedia-analytics

Simplistic architecture



How we get our data



How?

By Wikimedia Foundation / Joseph Allemandou ([1]) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Hive tables maintained

- Webrequest (raw and refined)
- pageview_hourly
- projectview_hourly
- pagecounts_all_sites
- mediacounts
- ...

Schemas: https://wikitech.wikimedia.org/wiki/Analytics/Data

Tables, schemas, and partitions

Live demo

- Accessing hive from stat1002
- Looking up tables and schemas
- Simple query
- Quick look at Hue

Privacy

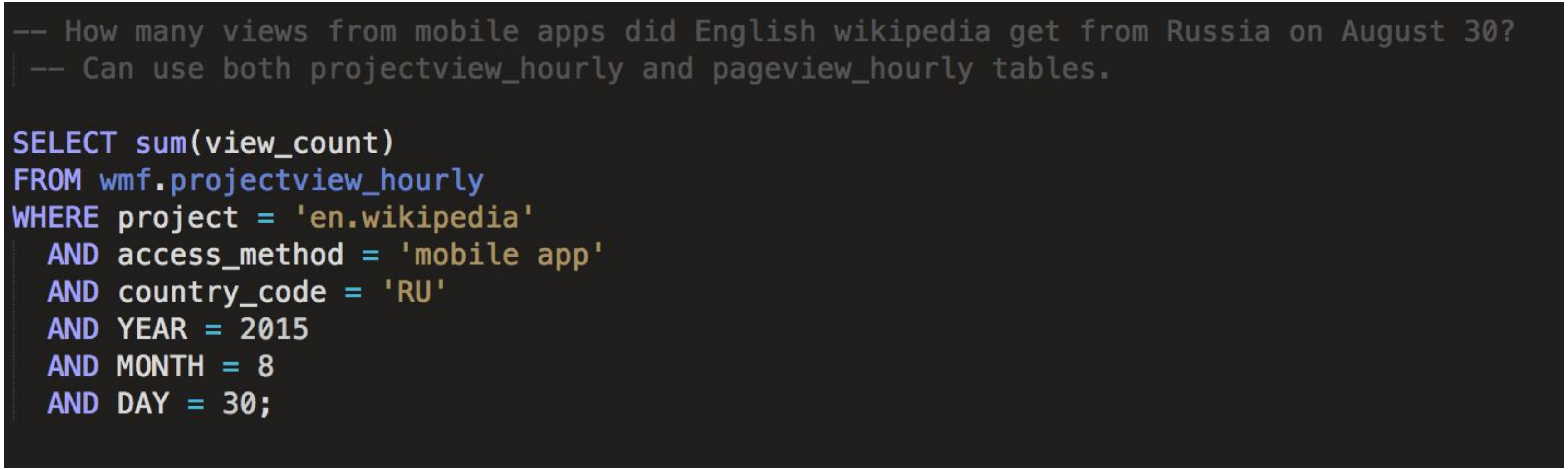
- Watch out for data that is potentially sensitive
- Keeping sensitive data private
- Special mentions - ips, user agents, Geolocation

https://wikimediafoundation.org/wiki/Privacy_policy

https://meta.wikimedia.org/wiki/Data_retention_guidelines

Let’s answer some questions

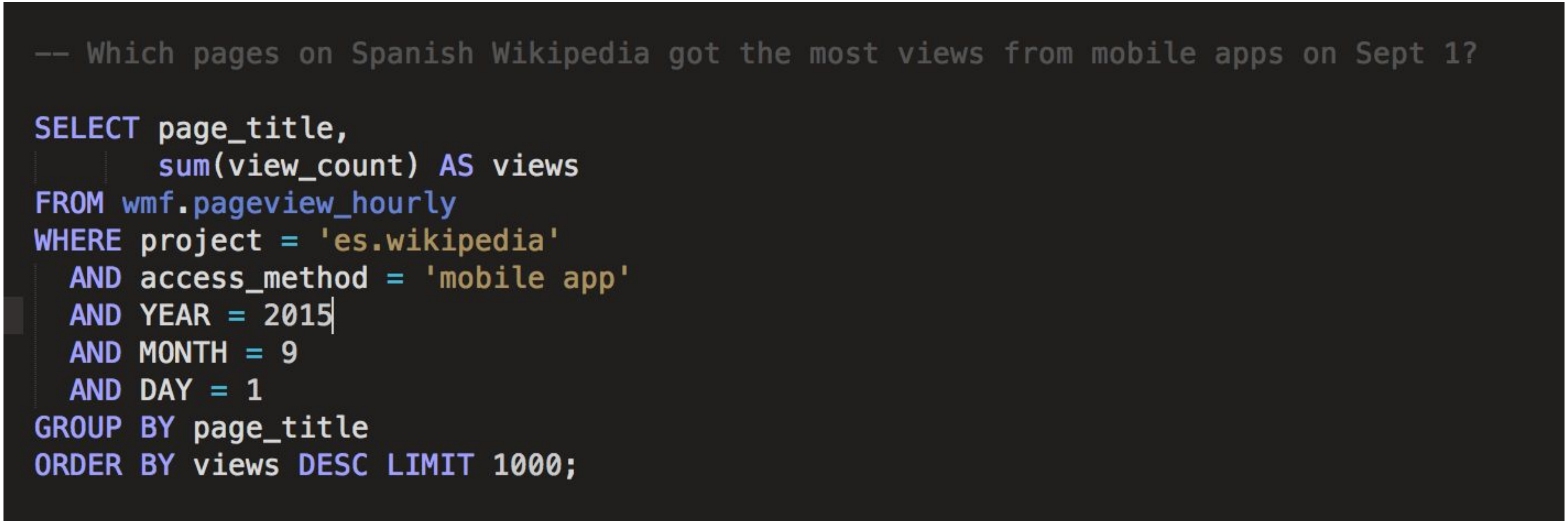
How many views from mobile apps did English wikipedia get from Russia on August 30?



```
-- How many views from mobile apps did English wikipedia get from Russia on August 30?
 -- Can use both projectview_hourly and pageview_hourly tables.

SELECT sum(view_count)
FROM wmf.projectview_hourly
WHERE project = 'en.wikipedia'
  AND access_method = 'mobile app'
  AND country_code = 'RU'
  AND YEAR = 2015
  AND MONTH = 8
  AND DAY = 30;
```

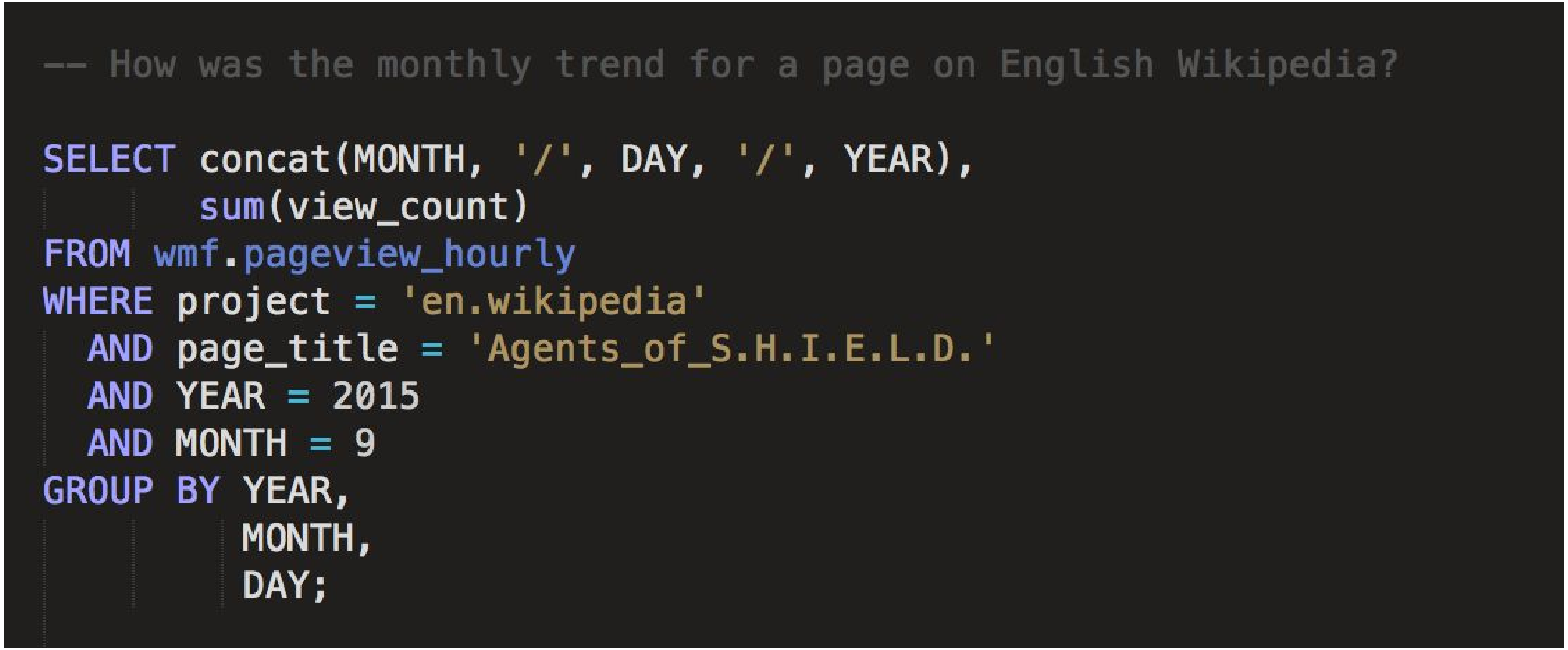
Which pages on Spanish Wikipedia got the most views from mobile apps on Sept 1?



```
-- Which pages on Spanish Wikipedia got the most views from mobile apps on Sept 1?

SELECT page_title,
       sum(view_count) AS views
FROM wmf.pageview_hourly
WHERE project = 'es.wikipedia'
  AND access_method = 'mobile app'
  AND YEAR = 2015
  AND MONTH = 9
  AND DAY = 1
GROUP BY page_title
ORDER BY views DESC LIMIT 1000;
```

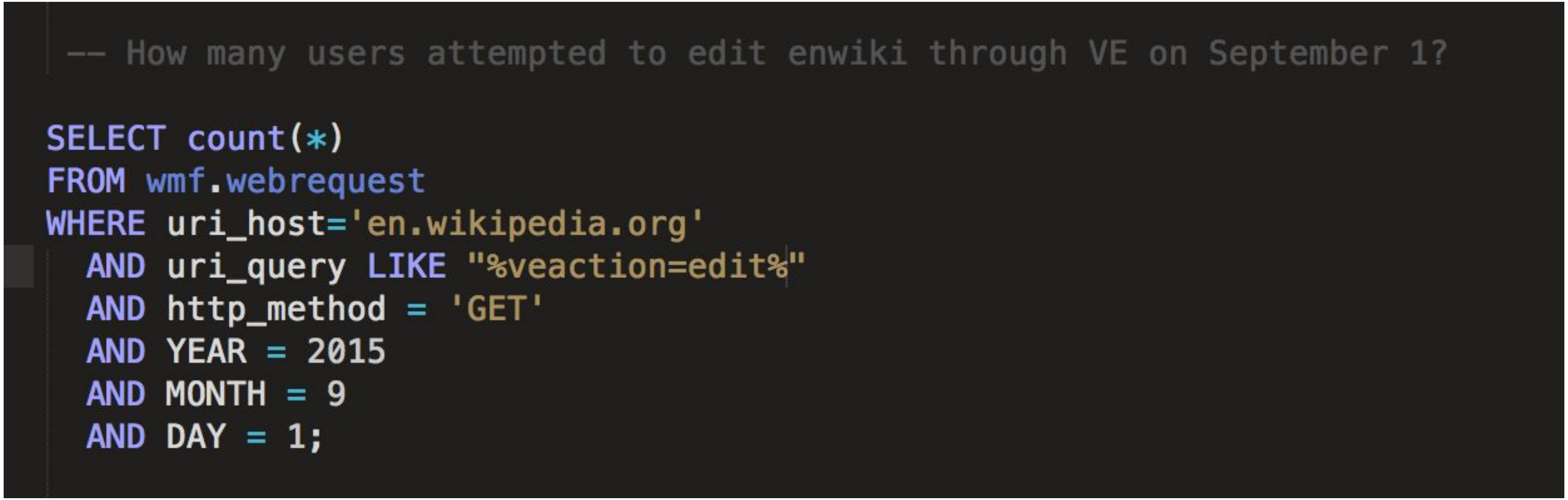
How was the monthly trend for a page on English Wikipedia?



```
-- How was the monthly trend for a page on English Wikipedia?

SELECT concat(MONTH, '/', DAY, '/', YEAR),
       sum(view_count)
FROM wmf.pageview_hourly
WHERE project = 'en.wikipedia'
  AND page_title = 'Agents_of_S.H.I.E.L.D.'
  AND YEAR = 2015
  AND MONTH = 9
GROUP BY YEAR,
         MONTH,
         DAY;
```

How many users attempted to edit enwiki through VE on September 1? Without VE?

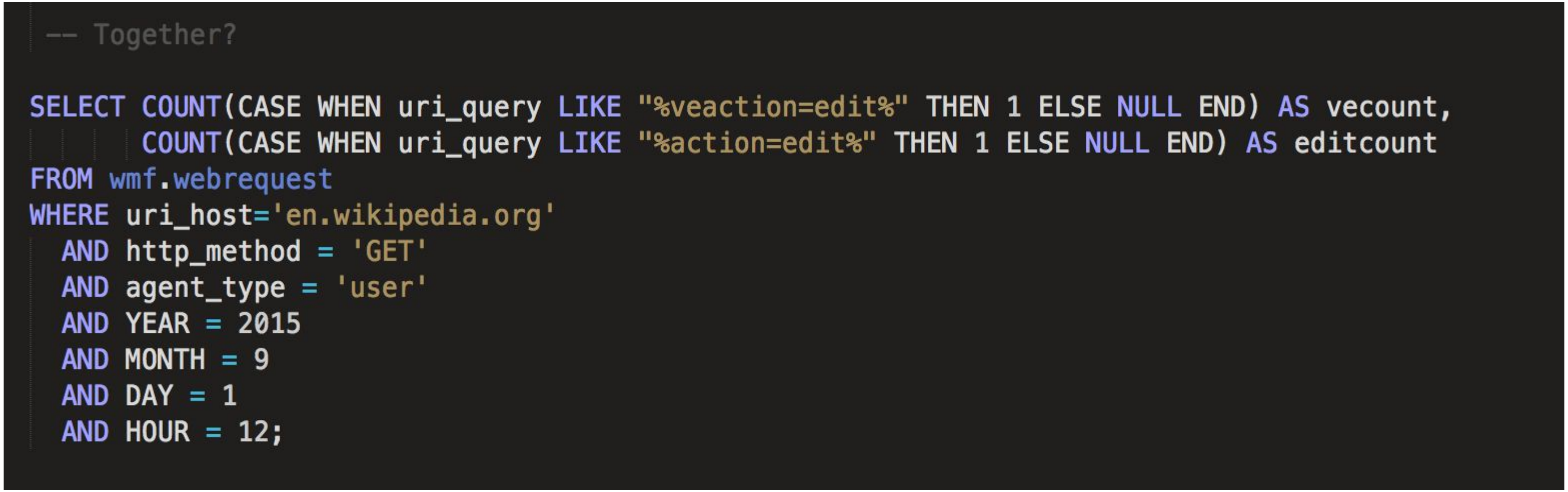


```
-- How many users attempted to edit enwiki through VE on September 1?

SELECT count(*)
FROM wmf.webrequest
WHERE uri_host='en.wikipedia.org'
  AND uri_query LIKE "%veaction=edit%"
  AND http_method = 'GET'
  AND YEAR = 2015
  AND MONTH = 9
  AND DAY = 1;
```

Note: This is just an example for uri_query and doesn't accurately the question - VE edit attempts cannot be measured by webrequests to pages with veaction=edit. The vast majority of ve edit attempts do not require an HTTP request of this kind. Rather they are initiated with JavaScript from the regular page view (which then virtually updates the browser url to include veaction=edit, but such request is never made, only in error scenarios and permalinks)

How many users attempted to edit enwiki through VE on September 1? Without VE?



```
-- Together?

SELECT COUNT(CASE WHEN uri_query LIKE "%veaction=edit%" THEN 1 ELSE NULL END) AS vecount,
       COUNT(CASE WHEN uri_query LIKE "%action=edit%" THEN 1 ELSE NULL END) AS editcount
FROM wmf.webrequest
WHERE uri_host='en.wikipedia.org'
  AND http_method = 'GET'
  AND agent_type = 'user'
  AND YEAR = 2015
  AND MONTH = 9
  AND DAY = 1
  AND HOUR = 12;
```

Note: This is just an example for uri_query and doesn't accurately the question - VE edit attempts cannot be measured by webrequests to pages with veaction=edit. The vast majority of ve edit attempts do not require an HTTP request of this kind. Rather they are initiated with JavaScript from the regular page view (which then virtually updates the browser url to include veaction=edit, but such request is never made, only in error scenarios and permalinks)

Tips and tricks

- Hive vs Hue
- Saving results to file
- Using screen for long running queries

Troubleshooting

- How to kill a running query?
- Out of memory errors
- Slow queries

https://wikitech.wikimedia.org/wiki/Analytics/Cluster/Hive/Queries#FAQ

Links

https://wikitech.wikimedia.org/wiki/Analytics/Cluster

https://wikitech.wikimedia.org/wiki/Analytics/Cluster/Hive

https://wikitech.wikimedia.org/wiki/Analytics/Cluster/Hive/Queries

https://hue.wikimedia.org/jobbrowser/ (Need access to analytics cluster)

http://yarn.wikimedia.org/cluster/scheduler (Need access to analytics cluster)

https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF

Questions?