Pageview

@

WIKIMEDIA
FOUNDATION

June 2015

Analytics Team

https://www.mediawiki.org/wiki/Analytics

Joseph Allemandou

joal@wikimedia.org - joal (IRC)

~ 9 billion HTTP hits per day

(2015 June 18th)

Make this data **useful**

Pageview

From Content Consumption Metrics page:

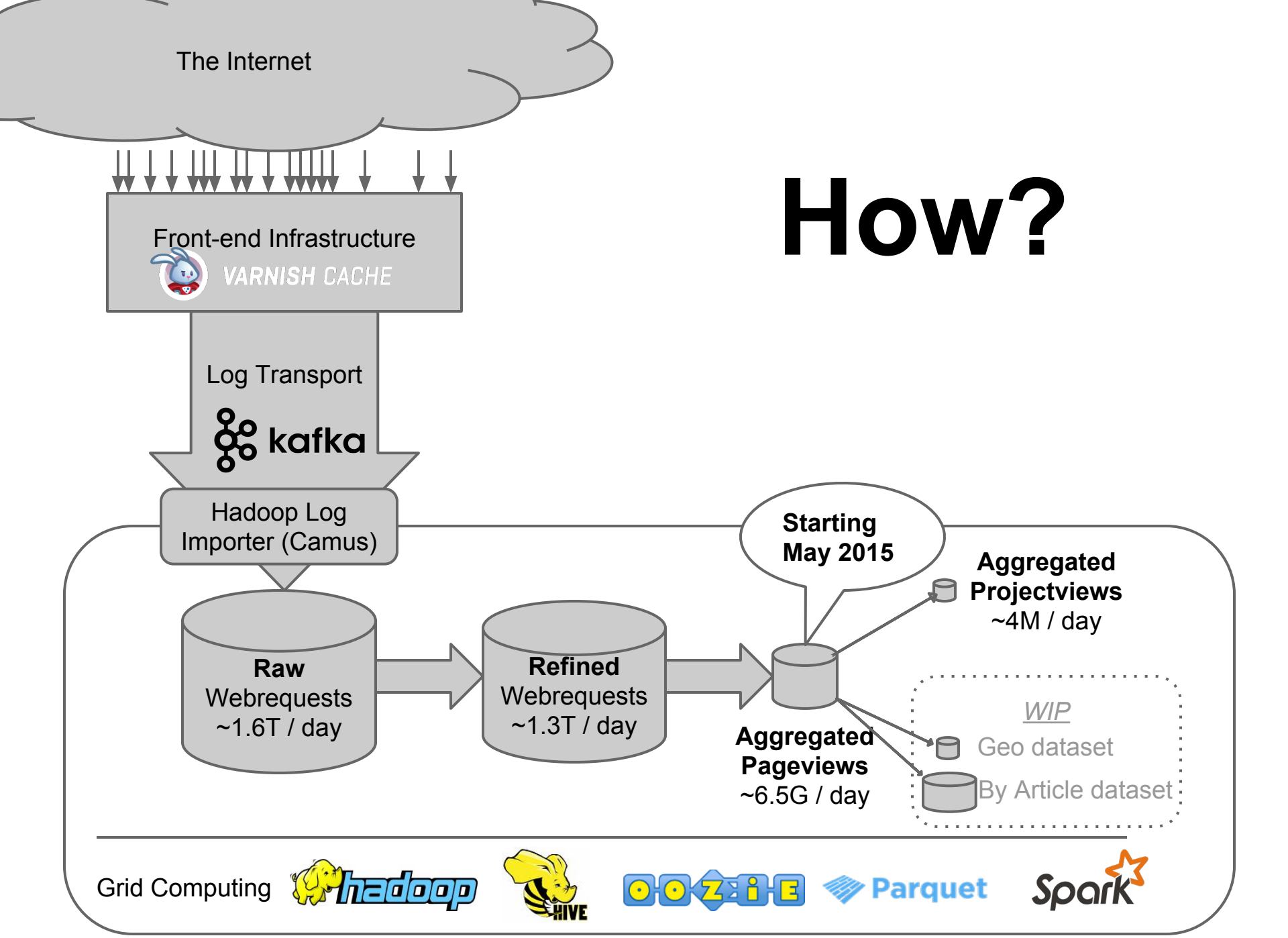
Page views are requests for Wikimedia assets that can be counted as *a single, human-driven request for a self-contained piece of text-based content*

What?

- **Unsampled** data
- **No spiders** (almost)
- **No actions (**except search**)**
- **Useful** additional **dimensions**

More info:
https://wikitech.wikimedia.org/wiki/Analytics/Pageviews/Definition
https://meta.wikimedia.org/wiki/Research:Page_view

How?


The Internet
Front-end Infrastructure
VARNISH CACHE
Log Transport
kafka
Hadoop Log Importer (Camus)
Raw
Webrequests
~1.6T / day
Refined
Webrequests
~1.3T / day
Starting May 2015
Aggregated Projectviews
~4M / day
Aggregated Pageviews
~6.5G / day
WIP
Geo dataset
By Article dataset
Grid Computing
hadoop
HIVE
OOZIE
Parquet
Spark

Who?

- For **now**
  - **internal** staff
  - Researchers under **NDA**

- In the **near future**
  - Everybody through an **API**
    Aggregated and **sanitized**

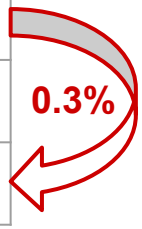
It does work!

pageviews - top 10 countries - desktop users - en.wikipedia - 2015-06-18

United States	79766449
Germany	26492802
Japan	20741594
Russia	16424765
United Kingdom	14945203
France	13529166
Canada	9146130
India	8751012
Italy	8654047
Brazil	7747792

Total Cumulative CPU time (comparable)

refined webrequests	~14 hours
aggregated pageviews	~16 minutes
aggregated projectviews	~3 minutes



WIP

- Dashboard integration
  - https://vital-signs.wmflabs.org

- **2 cubes** to deliver
  - **geo-oriented** dataset
  - **Page Title** dataset

- API to build
  - Reuse **RestBase**

Questions!

More details

- Refined Webrequests dataset
- Pre-aggregated pageviews dataset
- Pre-aggregated projectviews dataset