Scaling Wikidata Query Service

Split the Graph experiment

David Causse
Staff Software Engineer
(WMF)

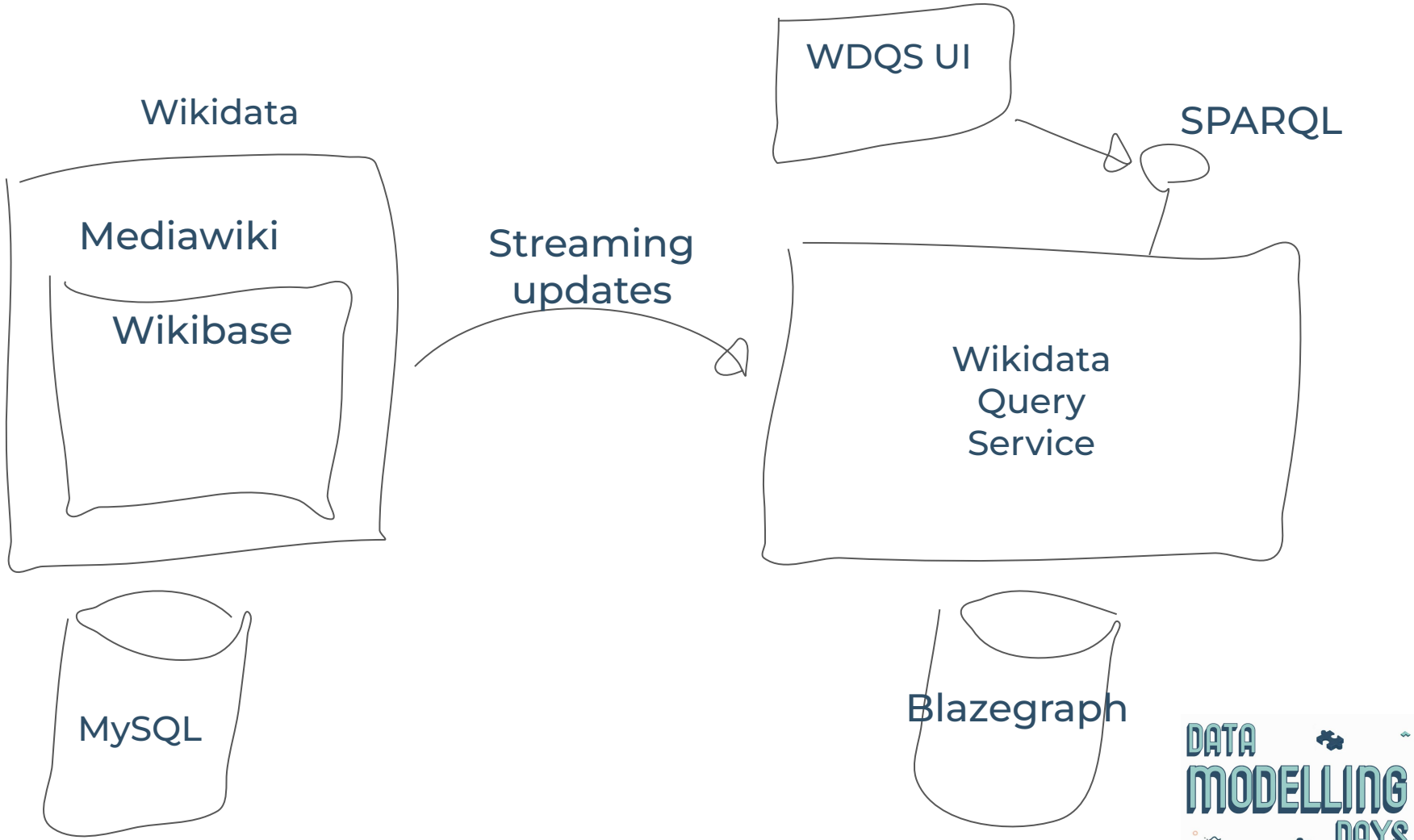
Guillaume Lederrey
Engineering Manager (WMF)

This session is recorded: Please mute your microphone and camera when you're not speaking.

DATA MODELLING DAYS
November 30
December 2
2023
WIKIDATA

Wikidata Query Service

DATA MODELLING DAYS
November 30
December 2
2023
WIKIDATA


WDQS UI
SPARQL
Wikidata
Mediawiki
Wikibase
Streaming updates
Wikidata Query Service
MySQL
Blazegraph
DATA MODELLING DAYS

WDQS use cases

- Edit / curation
- Visualizations (WDQS UI - https://query.wikidata.org/)
- Research
- Data reuse
- ...

Scaling challenges

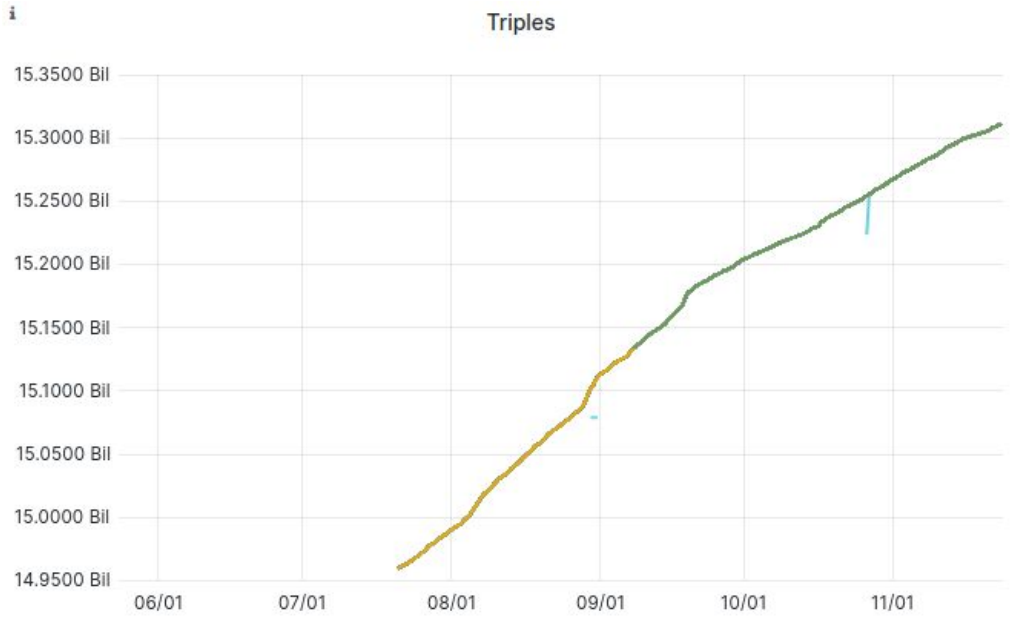
Scaling WDQS

3 facets

- Write load
- Read load
- Data size

Data Size

- ~15B triples
- Growing by ~1B triples/year
- One of the largest public SPARQL endpoints



DATA MODELLING DAYS

Data size limits

- Current backend (Blazegraph) has scaling issues
  - Random data corruption related to data size
  - Data reload taking ~ 1 month, with frequent failures
- Some queries are taking longer, and now reach timeout
- Not all alternative backends expose better scaling (https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_backend_update/WDQS_backend_alternatives)
- Scaling general purpose graphs is hard!

A dark future if we don’t act

- Without action, we might have a total failure of WDQS
  - https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/WDQS_backend_update/Blazegraph_failure_playbook
  - Unable to ingest new edits
  - Unable to recover from failure (data reload from dumps takes > 1 month, when it works)

- We’re not there yet!
  - https://grafana.wikimedia.org/d/slo-WDQS/wdqs-slo-s?orgId=1

Potential options

Options

- Reduce the size of Wikidata
  - Reduce duplication
    - MUL
    - Automated descriptions
  - Let’s not go there
    - Stop editing
    - Delete data
- Split the Graph and use Federation
  - Federation is core to Linked Open Data and already happening (~100 SPARQL endpoints in the WDQS federation allow list)
  - Provides a way to scale horizontally

Constraints of a split

- Easy to understand and navigate
- The largest component is at most 75% of the current graph size
- Minimize the number of queries needing rewriting due to federation
- Minimize the number of queries rendered too expensive by federation
- Query time is not increased for most queries

Potential split

- Truthy vs reified
- Labels, descriptions, aliases, ...
- Topical split:
  - Astronomical objects
  - Scholarly articles

Scholarly articles

- ~50% of the graph
  - https://wikitech.wikimedia.org/wiki/User:AKhatun/Wikidata_Subgraph_Analysis#Table_of_top_50_subgraph_information
- ~2.5% of queries
  - https://wikitech.wikimedia.org/wiki/User:AKhatun/Wikidata_Subgraph_Query_Analysis#Query_count_and_time
- Easy to understand (instance of Q13442814)

What is the plan

DATA MODELLING DAYS
November 30
December 2
2023
WIKIDATA

Scholarly article split in WDQS

- Expose 3 experimental endpoint:
  - Full graph
  - Scholarly article
  - Main graph (Full graph - scholarly articles)
- Test on a significant number of queries (from logs)
- Gather feedback from you all

Success criteria

- Stability of our backend isn't threatened by the size of the graph
    - Wikidata knowledge graph can be reloaded within 10 days for a graph of up to 20 billion tuples
- Query time does not increase for most queries
- Number of queries requiring rewrite due to federation is minimal
- Number of queries rendered too expensive by federation is minimal

What we’re not doing

- No changes to Wikidata, only to WDQS
- No other experiment beside Scholarly Article
  - We might explore other split if this experiment isn’t successful
- No rollout before validating this is a viable solution
  - Implementation will take time!

Timeline (optimistic)

- By end of January 2024
  - availability of a somewhat stable testing environment
- By end of January 2024
  - testing of the split on a subset of existing queries, feedback from all of you about how this split is functioning for different workloads
- By end of March 2024
  - reflection on this experiment and next steps, either experimentation with a different split, or start to productionize the current one

Thanks for your attention!

Get in touch with us:

David Causse
dcausse@wikimedia.org

Guillaume Lederrey
gehel@wikimedia.org

Search Platform team
https://wikitech.wikimedia.org/wiki/Search_Platform/Contact

Credits
Information about the licence of your slides, the pictures used in your slides, links to additional resources, etc.

DATA MODELLING DAYS
November 30
December 2
2023
WIKIDATA