Incremental update of data lineage storage in a graph database

Jan Sýkora | Supervisor: Ing. Michal Valenta, Ph.D. | Czech Technical University in Prague | Faculty of Information Technology

Problem statement

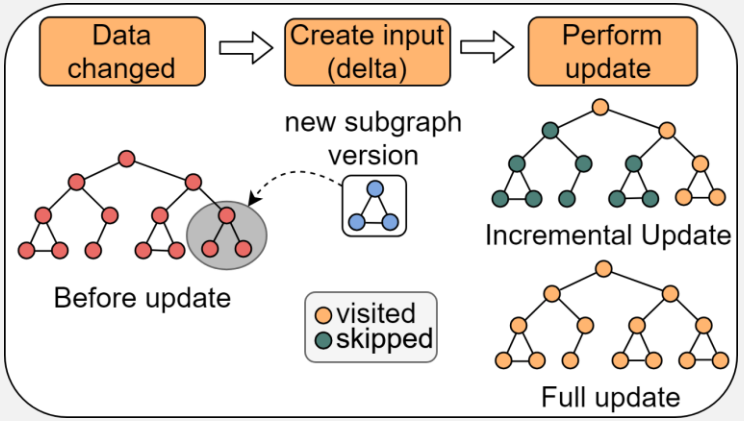
- This thesis addresses an issue of ineffective updates of data lineage storage in a graph database Titan used in tool Manta Flow.
- The time spent on updating a change was directly proportional to the overall size of all data, regardless the change size - only Full update was possible.

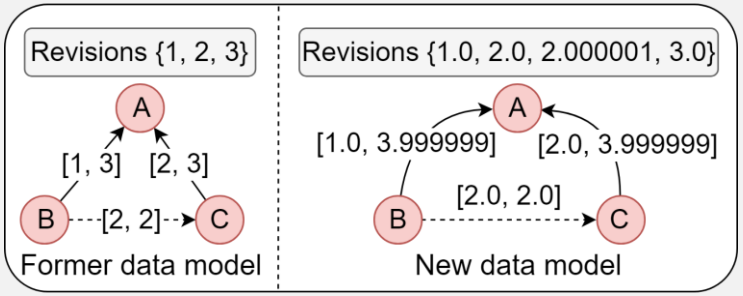
Goal of the thesis

- The main goal of this thesis was to design and implement Incremental update of the data lineage storage → update only the changed data and avoid unnecessary updates of the unchanged data.

Improved data model

- Data model is a **tree structure**
- Each edge and vertex has its revision validity stored on an edge
- Former data model uses closed end revisions → impossible to update only a part of the graph
- New data model introduces subrevisions and new representation of end revisions
- New data model allows to update only a changed part of the graph





Update type \ Database size / #deltas	Small DB (20 deltas)	Big DB (145 deltas)	Large DB (29 deltas)
Full Update *(former data model)*	78 930	197 771	373 503
Full Update *(new data model)*	89 870	218 181	379 400
Incremental Update *(new data model)*	1 918	13 453	2 790

Table: Merge time comparison [ms]

New update method

1. Create input subgraph representing the changed data (delta)
2. Merge input subgraph to the database graph starting from the root
3. When the corresponding subgraph in the database is reached during merging, its latest version is stored in history and the new subgraph version is added

Performance testing

- Small changes: Incremental update is significantly faster than Full update
- Large changes: Incremental update has a similar performance as Full update
- Full update time performance on the new data model is similar to Full update time performance on the former data model

Results

- Incremental update is very effective and allows fast updates (esp. small changes)
- New data model and update method is already deployed in the product version of software tool Manta Flow
- Results of the thesis are applicable in another graph database oriented apps

FACULTY OF INFORMATION TECHNOLOGY
CZECH TECHNICAL UNIVERSITY IN PRAGUE