LDBC SNB Interactive 2.0

David Püroja, Gábor Szárnyas, Jack Waudby, Peter Boncz

15th LDBC Technical User Community meeting

Comparison of workloads

	Interactive v1.0	Business Intelligence v1.0
focus	OLTP	OLAP
typical query	2-3 hop neighbourhood queries with filtering	multi-hop/path/subgraph queries with filtering & aggregation
data generator	SNB Hadoop Datagen	SNB Spark Datagen
refresh operations	inserts	inserts and deletes
target metric	Throughput (ops/s)	mean latency (WIP)

Interactive v1.0 and v2.0

	Interactive v1.0	Interactive v2.0
focus	OLTP	OLTP
typical query	2-3 hop neighbourhood queries with filtering	2-3 hop neighbourhood queries with filtering
data generator	SNB Hadoop Datagen	SNB Spark Datagen
refresh operations	**inserts**	**inserts and deletes**
target metric	Throughput (ops/s)	Throughput (ops/s)

Interactive Workload

Interactive workload

Scenario: Users browsing a social network and producing content (Forums, Messages) and *delete content or their account*

Queries:

- 14 complex reads
- 7 short reads
- 8 insert operations
- **8 delete operations**

Queries

- **Complex queries:** Always start from one or two Person nodes, and discover their neighbourhoods (1..2 nodes) or paths between Person nodes

- **Short queries:** Discover the neighbourhood of a Person or a Message node

- **Insert operations:** Each operation inserts a node (and connects it to its neighbourhood) or an edge between existing nodes

- **Delete operations:** Each operation deletes a node (and its connections to its neighbourhood) or an edge between existing nodes *(new in v2.0)*

Interactive deletions

Deletions are backported from BI workload. The delete queries are:

1. Remove person and its personal forums and message (sub)threads
2. Remove post like
3. Remove comment like
4. Remove forum and its content
5. Remove forum membership
6. Remove post thread
7. Remove comment subthread
8. Remove friendship

Execution of queries

- **Insert and delete queries:** operations issue times are taken from the update streams generated by the data generator

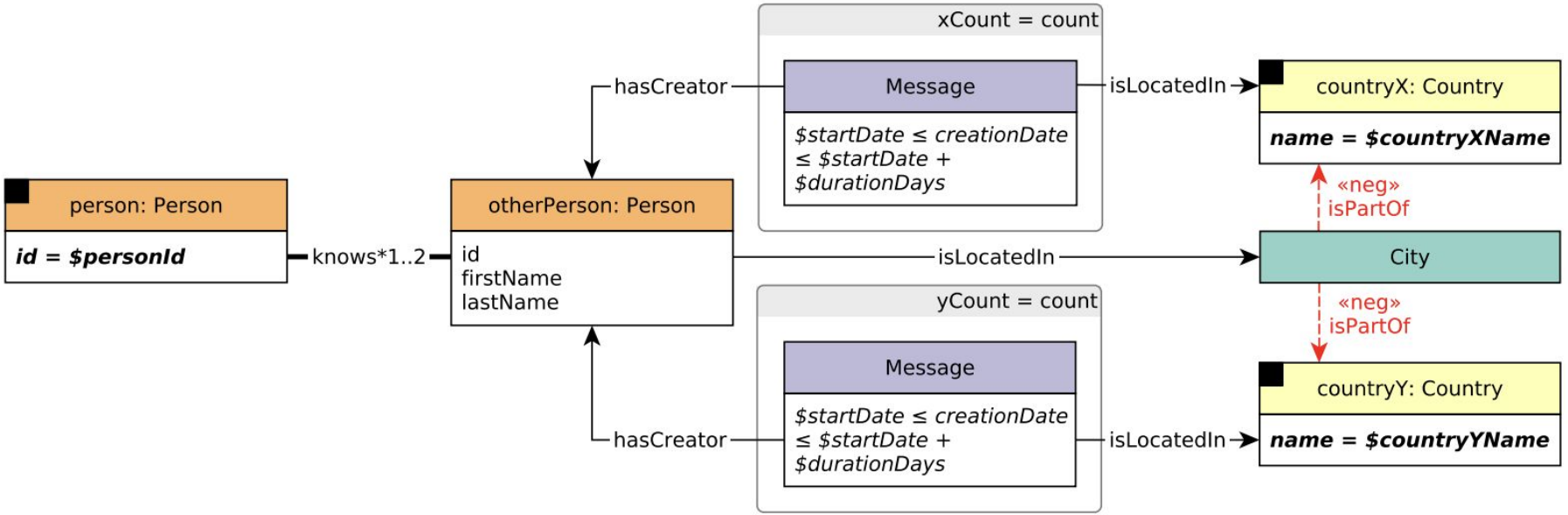
- **Complex queries:** complex reads times are expressed in terms of update operations (update frequencies)

- **Short read queries:** for each complex read instance, a sequence of short reads is planned

Example queries

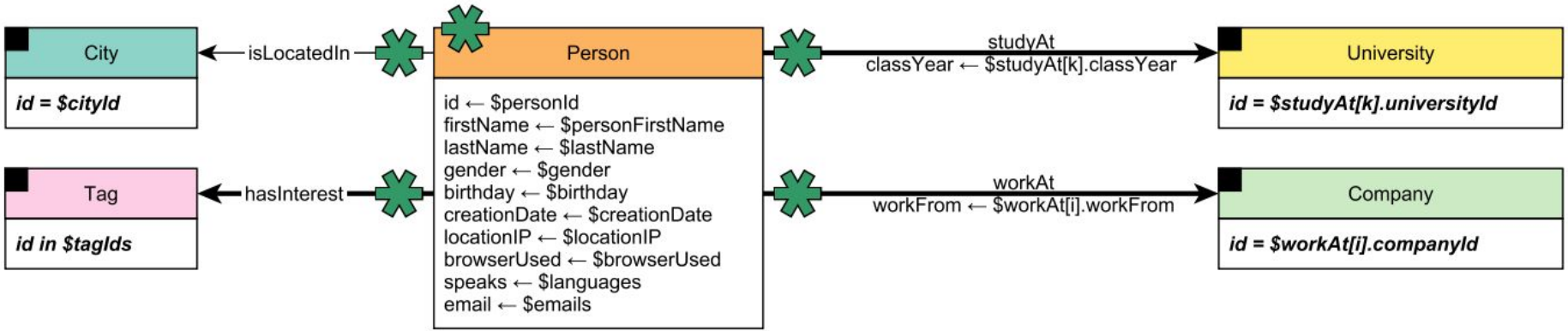
Example queries: Complex query 3

Q3: Friends and friends of friends that have been to given countries



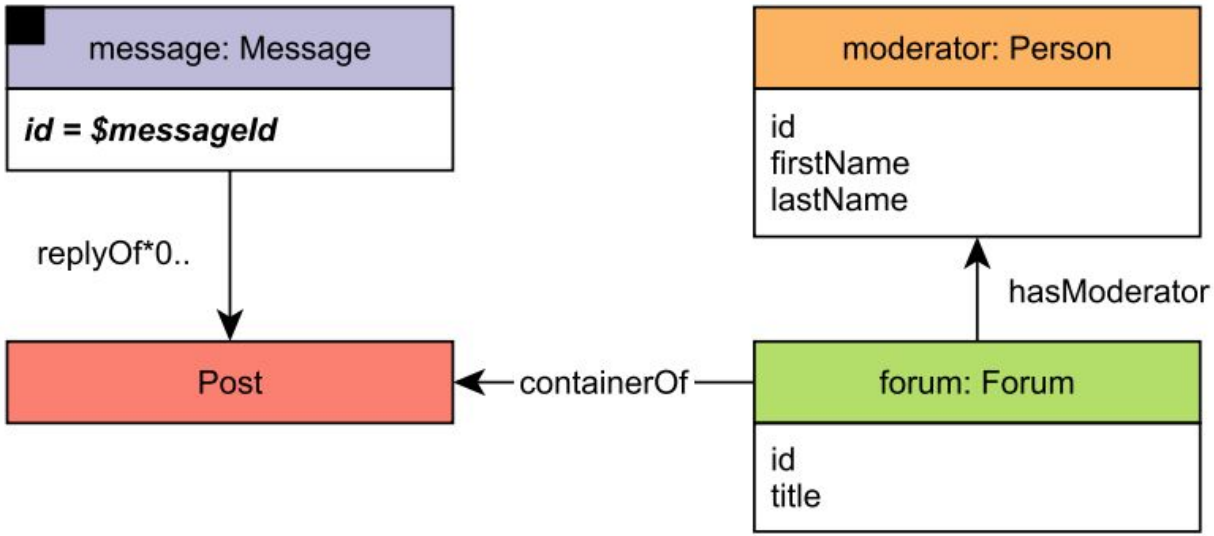
Example queries: Insert query 1

Q1: Add person



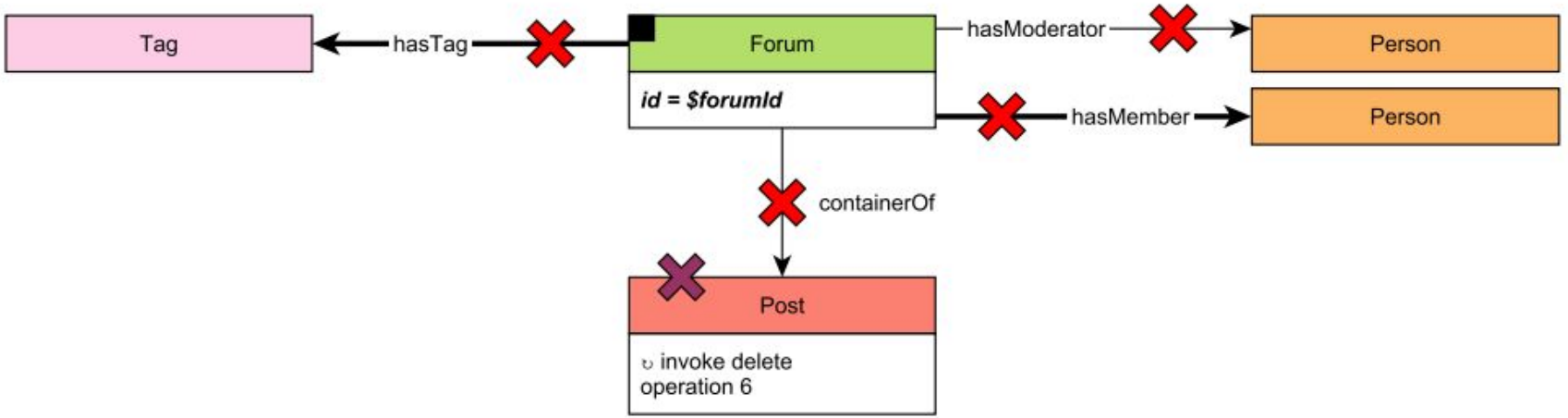
Example queries: Short query 6

Q6: Forum of a message



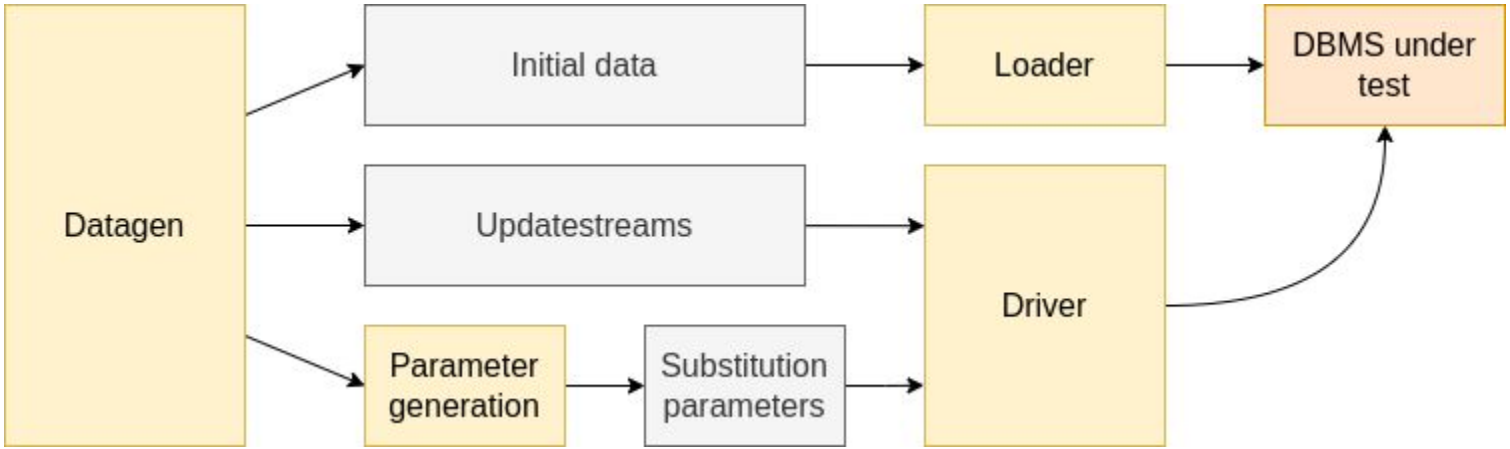
Example queries: Delete query 4

Q4: Remove forum and its content



Interactive Driver

Interactive Driver: Components



Interactive Driver: Implementations

- PostgreSQL [SQL]: a row-oriented RDBMS
- Neo4j [Cypher]: a graph database management system
- DuckDB [SQL]: a column-oriented OLAP RDBMS with a vectorized runtime
- Umbra [SQL]: a column-oriented HTAP RDBMS with a compiled runtime, WIP
- TigerGraph [GSQL]: graph database management system
- SQL Server [T-SQL]: row or column-oriented RDBMS with graph extension, WIP

Interactive Driver: v1.0 changes

- Scale Factor properties included in driver
- Forum/Person update streams merged
- Driver divides the update stream across update-threads round-robin
- Fix 5% on-time validation rule
- Short reads are separated from update queries

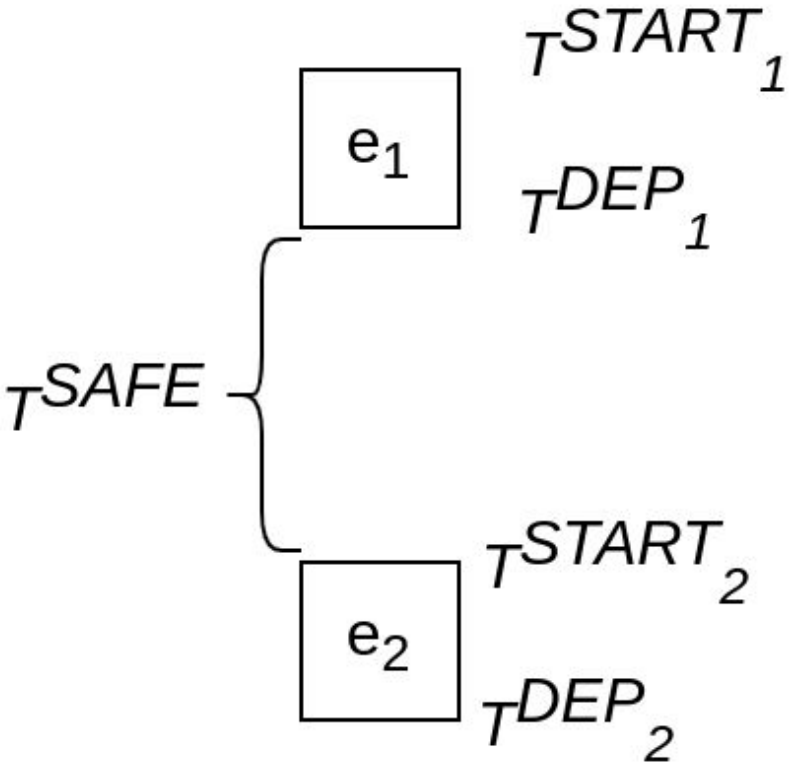
Interactive Driver: v2.0 Future work

- Move to dataset from Spark Datagen
  - Allows **higher scale factors**
  - Include **deletions**
- Implement **windowed execution**[1] to support asynchronous execution of dependent events
- Improve speed of parameter generation
- Scoring analysis separate from driver (using DuckDB)

[1] *Orri Erling et al. "The LDBC social network benchmark: Interactive workload." SIGMOD 2015.*

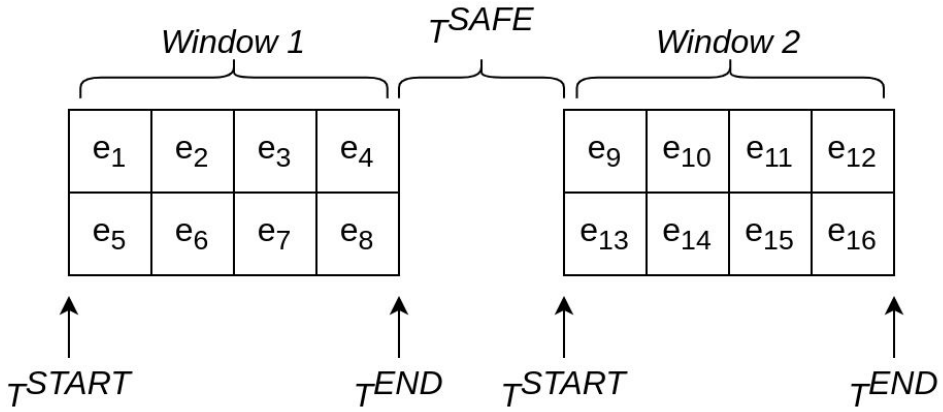
Windowed Execution

1. Event 1 has a starttime and a dependent time
2. Event 2 is dependent on event 1
3. Datagen ensures that there is T-safe time between the T-dep time of event 1 and the T-start time of event 2



Windowed Execution

When taking multiple events into account, each starting in a set interval, the window with dependent events is scheduled T-safe time



Larger scale factors

- Current Interactive workload goes up to SF1000
- New scale factors are:
  - SF3000
  - SF10k
  - SF30k (Work in progress)

Interactive workload: Audited results

4 audited results:

1. Sparksee (driver v0.2.2, 2015)
2. Virtuoso (driver v0.2.2, 2015)
3. TuGraph (driver v0.3.2, 2020) (report)
4. CreateLink (driver v0.3.3, 2022) (report)


LDBC
The graph & RDF
benchmark reference