The LDBC benchmark suite

Gábor Szárnyas

(LDBC benchmark expert)

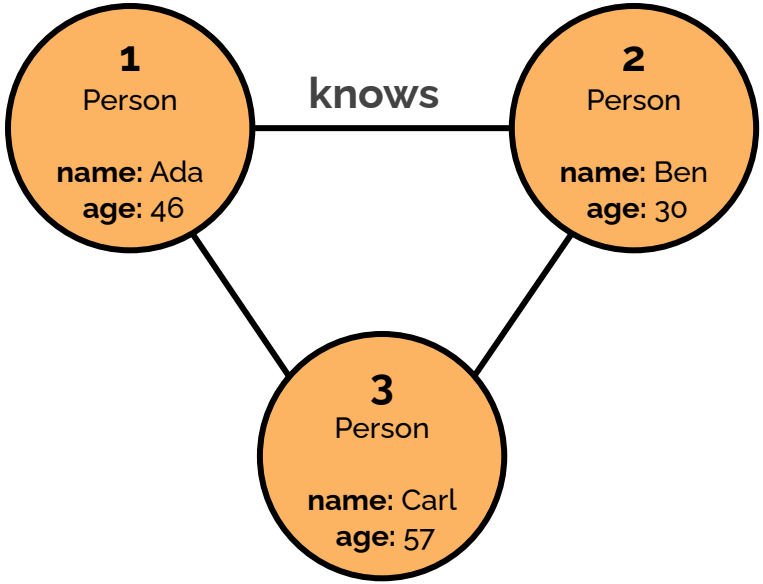
Data modelling: Tabular vs. graph

Person

id	name	age
1	Ada	46
2	Ben	30
3	Carl	57

knows

person a	person b
1	2
2	3
1	3



Waves of the “attributed graph” data model

year	data model	declarative language
1969	network model (CODASYL)	no
1988	object-oriented model	OQL
1999	RDF	SPARQL
2010	property graph	Cypher, Gremlin, …

Graph databases (2010-)

```
MATCH
  (p1:Person)-[:knows]-(p2)
  (p2:Person)-[:knows]-(p3)
  (p3:Person)-[:knows]-(p1)
```

pattern matching

```
MATCH

(p1:Person)-[:knows*]-(p2:Person)
```

path-finding

STAMFORD, Conn., March 16, 2021

Gartner Identifies Top 10 Data and Analytics Technology Trends for 2021

The (sorry) State of Graph Database Systems

Peter Boncz
CWI

The Register

The Great Graph Debate: Revolutionary concept in databases or niche curiosity?

Knowledge graphs 'overcome the shortcomings of large language models'

Investing in knowledge graphs provides higher accuracy for LLM-powered, question-answer systems over SQL databases, data.world's Juan Sequeda, says

02 Feb 2024 | INTERVIEWS

Waves of the “attributed graph” data model

year	data model	declarative language
1969	network model (CODASYL)	no
1988	object-oriented model	OQL
1999	RDF	SPARQL
2010	property graph	Cypher, Gremlin, …

problem #1:
usually no standard query language

problem #2:
performance limitations

Competition drives performance!

Initially, RDBMSs also had serious performance problems

1980s: *benchmark wars*

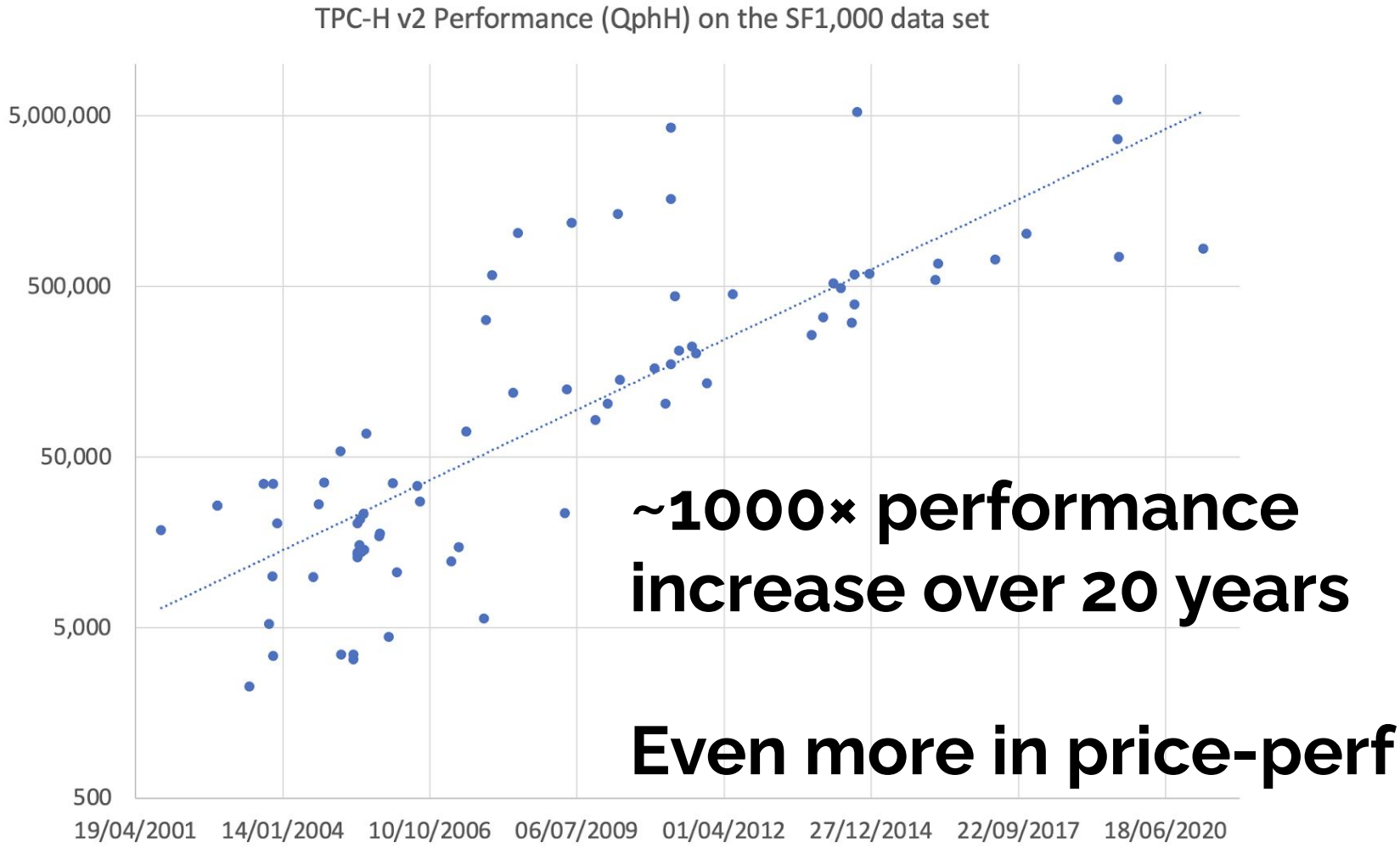
- Objective system-to-system comparison is very difficult
- Vendors are motivated to boast good results
- Need an independent authority and a standard

Inspiration: TPC benchmarks

TPC®

Transaction Processing Performance Council (1988–)

Influential benchmarks: TPC-C, TPC-H, TPC-DS


TPC-H v2 Performance (QphH) on the SF1,000 data set
5,000,000
500,000
50,000
5,000
500
19/04/2001
14/01/2004
10/10/2006
06/07/2009
01/04/2012
27/12/2014
22/09/2017
18/06/2020
~1000× performance increase over 20 years
Even more in price-perf

LDBC: Linked Data Benchmark Council

A non-profit company

~25 organizational and 100 individual members

Mission: Accelerate progress in graph data management

 ldbcouncil.org

 github.com/ldbc

Stakeholders

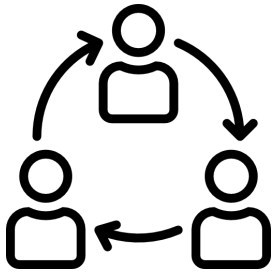
database companies

hardware vendors

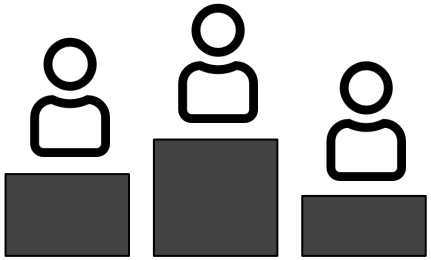
cloud providers

researchers and academic institutes

LDBC encourages stakeholders to...



collaborate on standards



compete on performance

Sponsor Companies

ANT GROUP

火山引擎 Beijing Volcano (ByteDance)

ORACLE Labs

Companies and Research Institutes

Alibaba Cloud

aws

intel

neo4j

TigerGraph

JCC Consulting, Inc.

MEM GRAPH

ontotext

fabarta

*Sparsity

创邻科技 CREATE LINK

pometry

ULTIPA

ArangoDB

海致星图 StarGraph

NebulaGraph

RelationalAI

thatDot

Birkbeck UNIVERSITY OF LONDON

FORTH

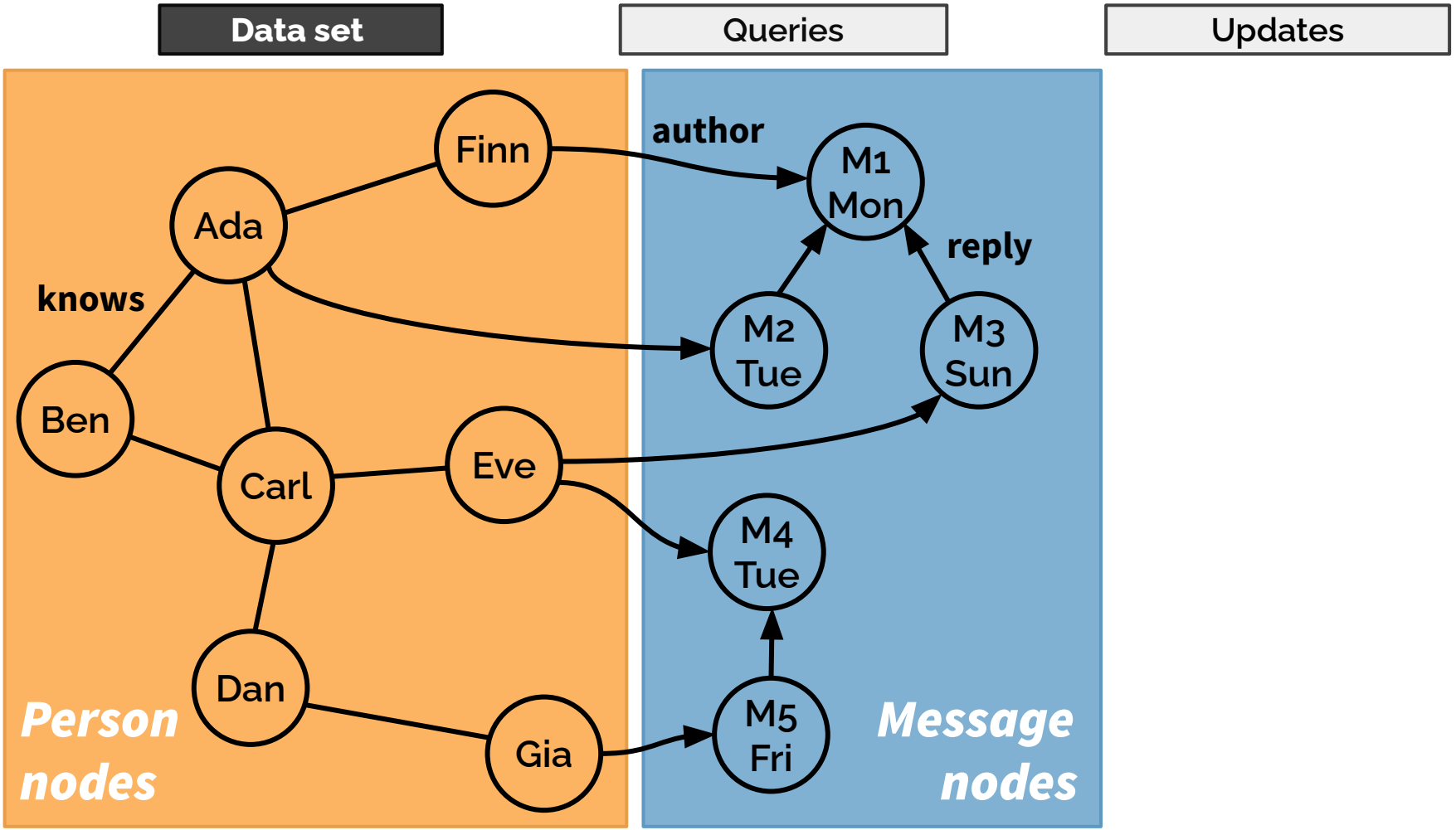
Database workloads

Social Network Benchmark suite

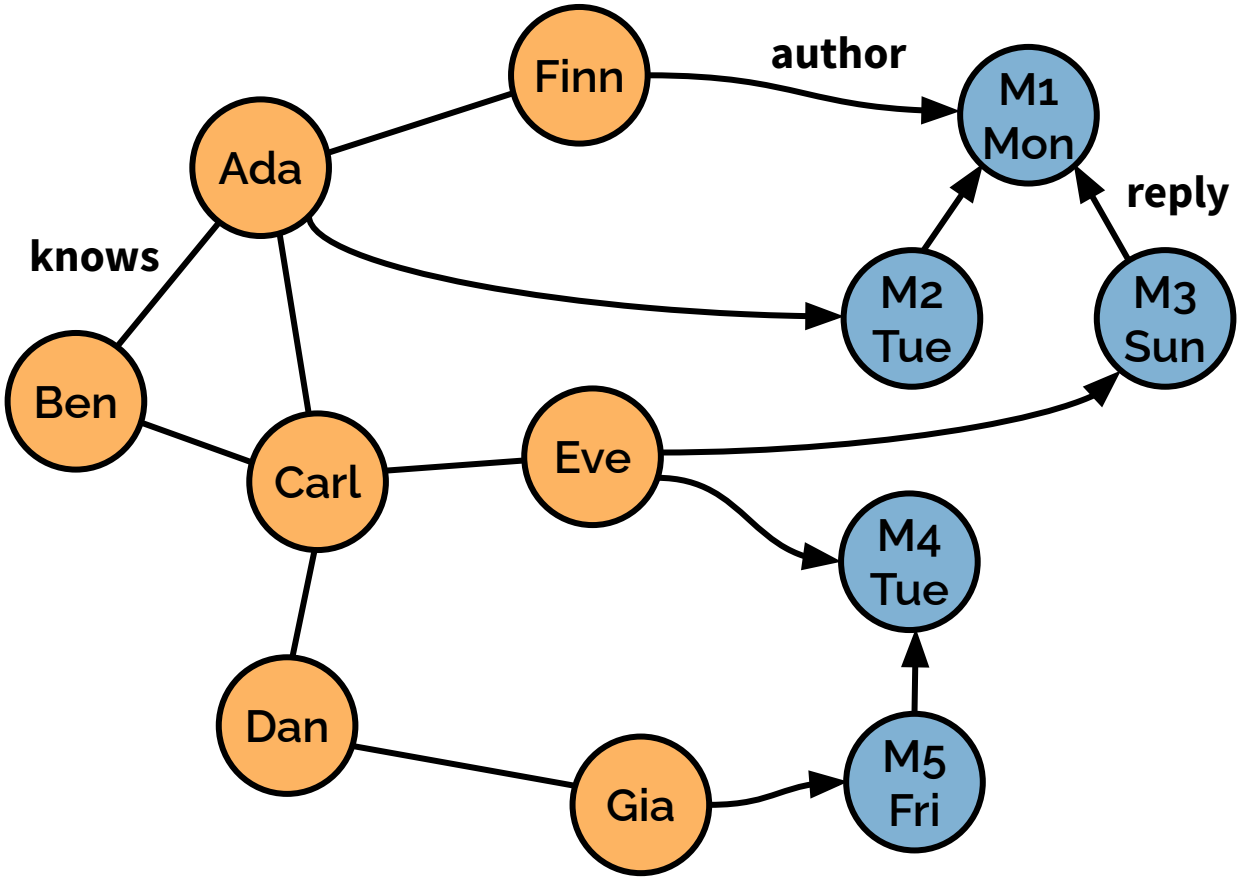
Data set

Queries

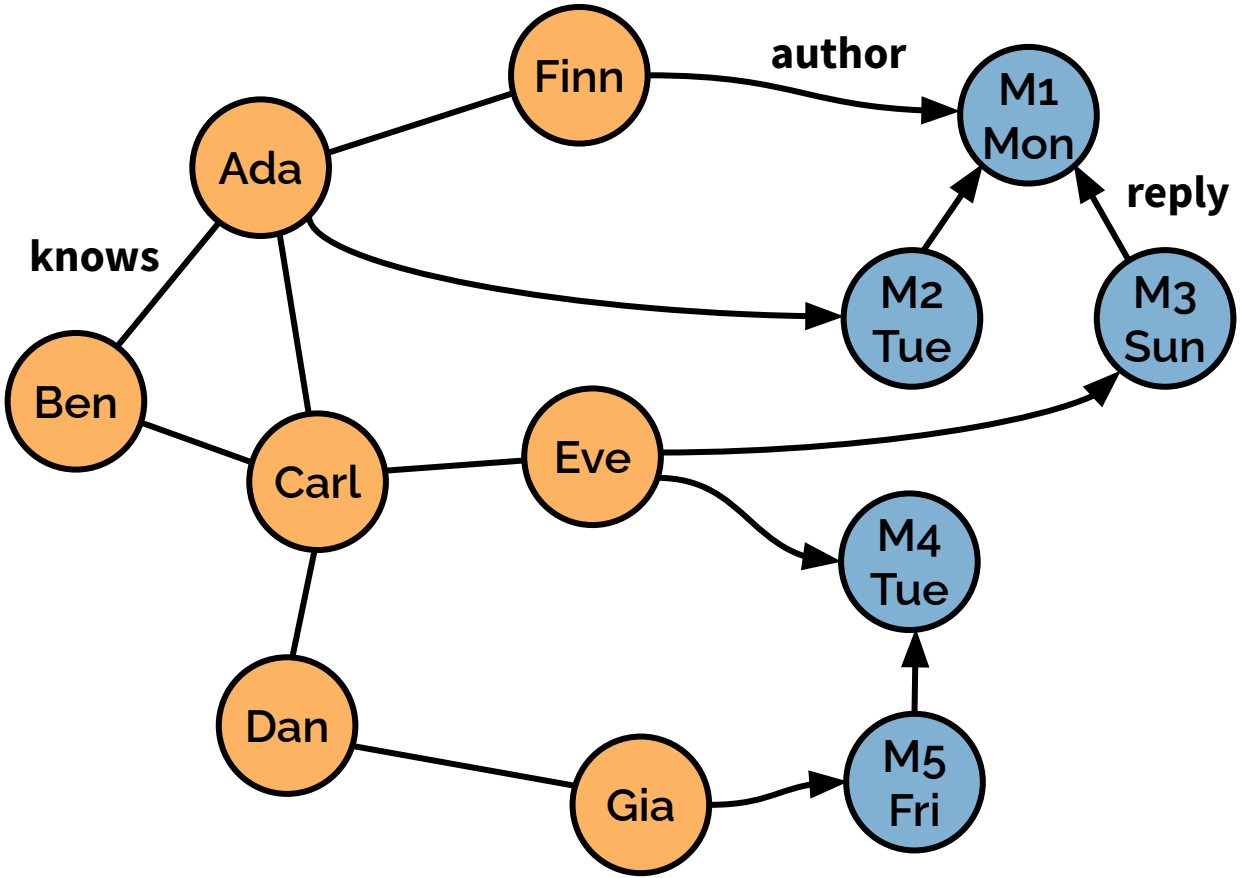
Updates


Data set
Queries
Updates
Finn
author
M1
Mon
Ada
reply
knows
M2
Tue
M3
Sun
Ben
Carl
Eve
M4
Tue
Dan
Person
nodes
Gia
M5
Fri
Message
nodes

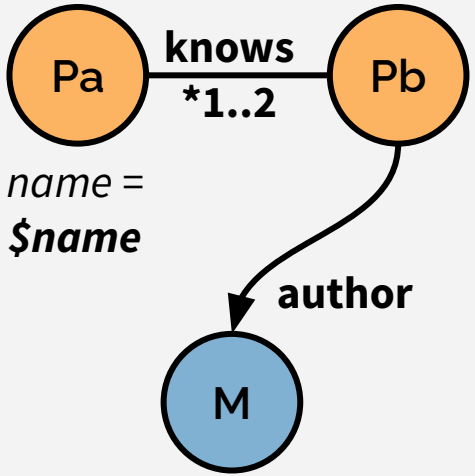
Data set | **Queries** | Updates



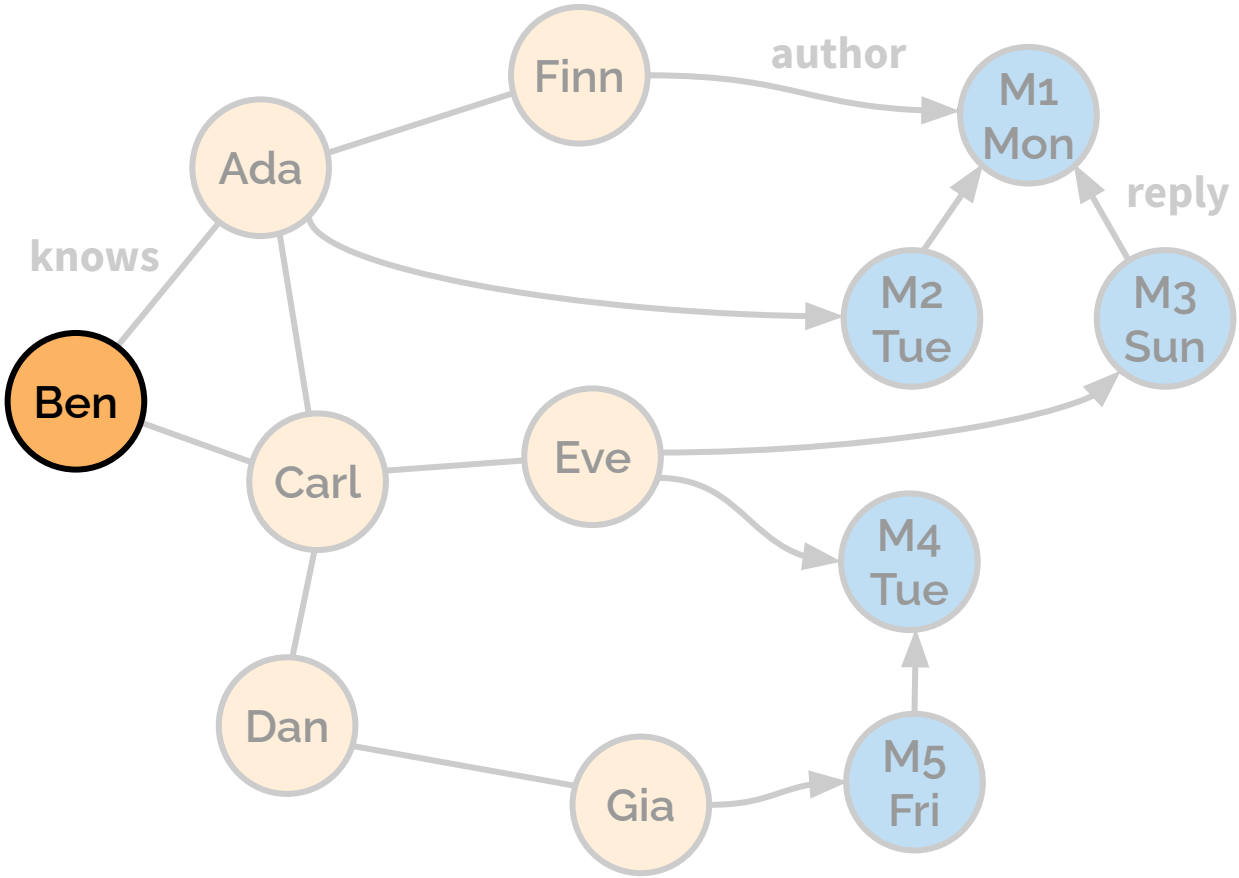
Data set | **Queries** | Updates



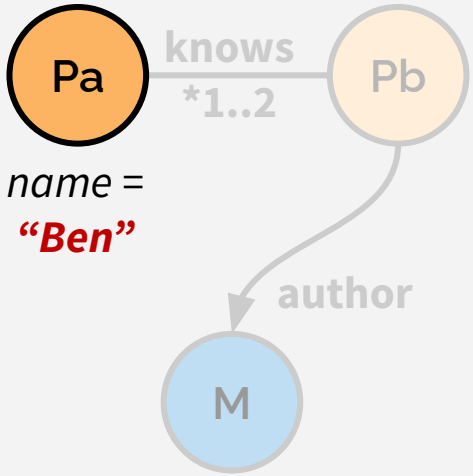
Q9(*$name*, *$day*)



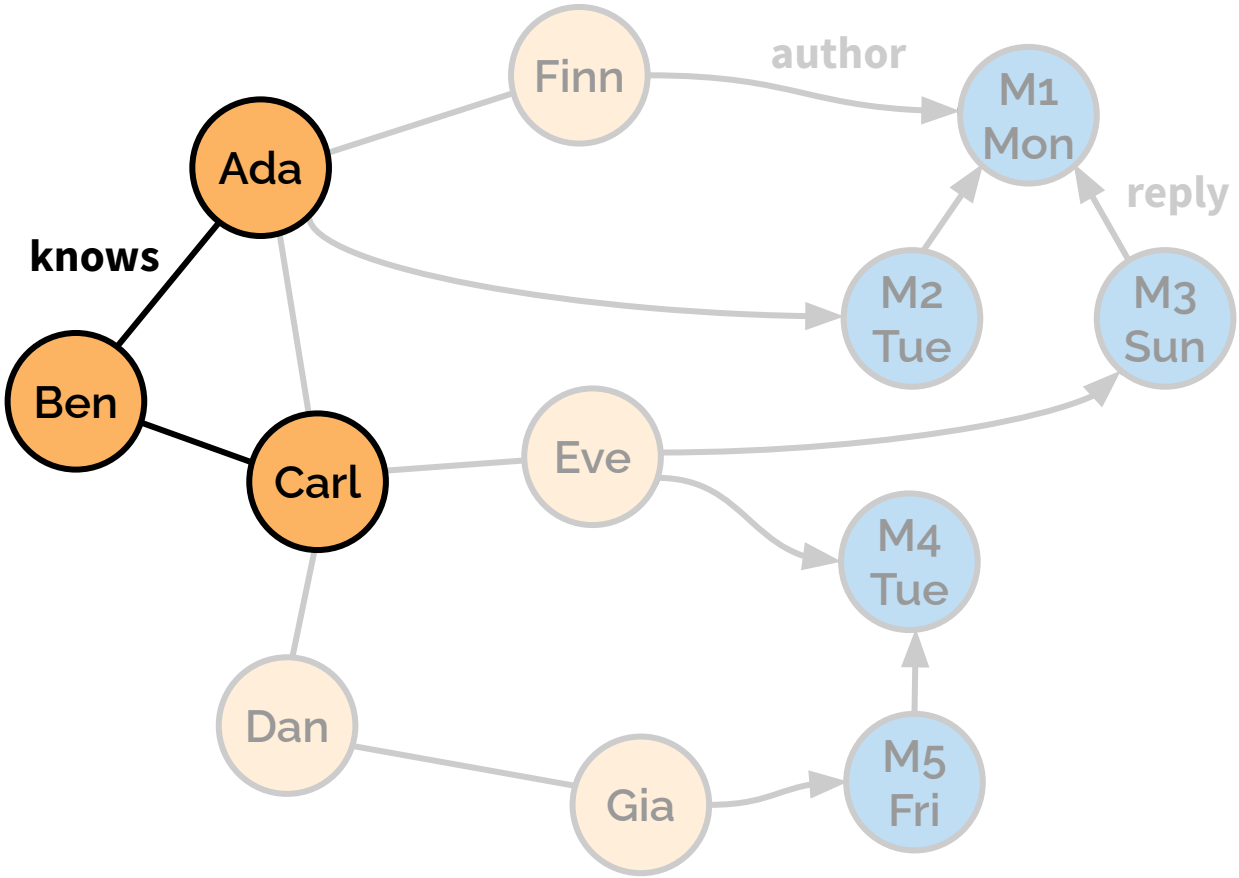
Data set | **Queries** | Updates



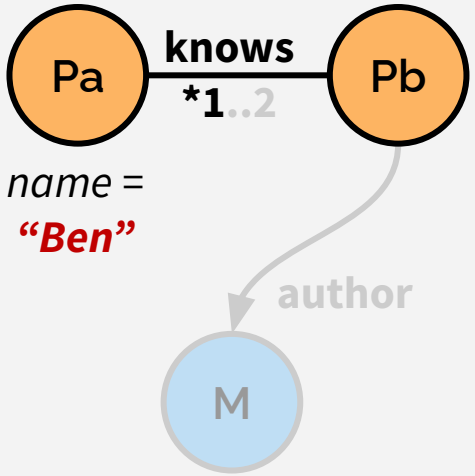
Q9(*"Ben"*, *"Sat"*)



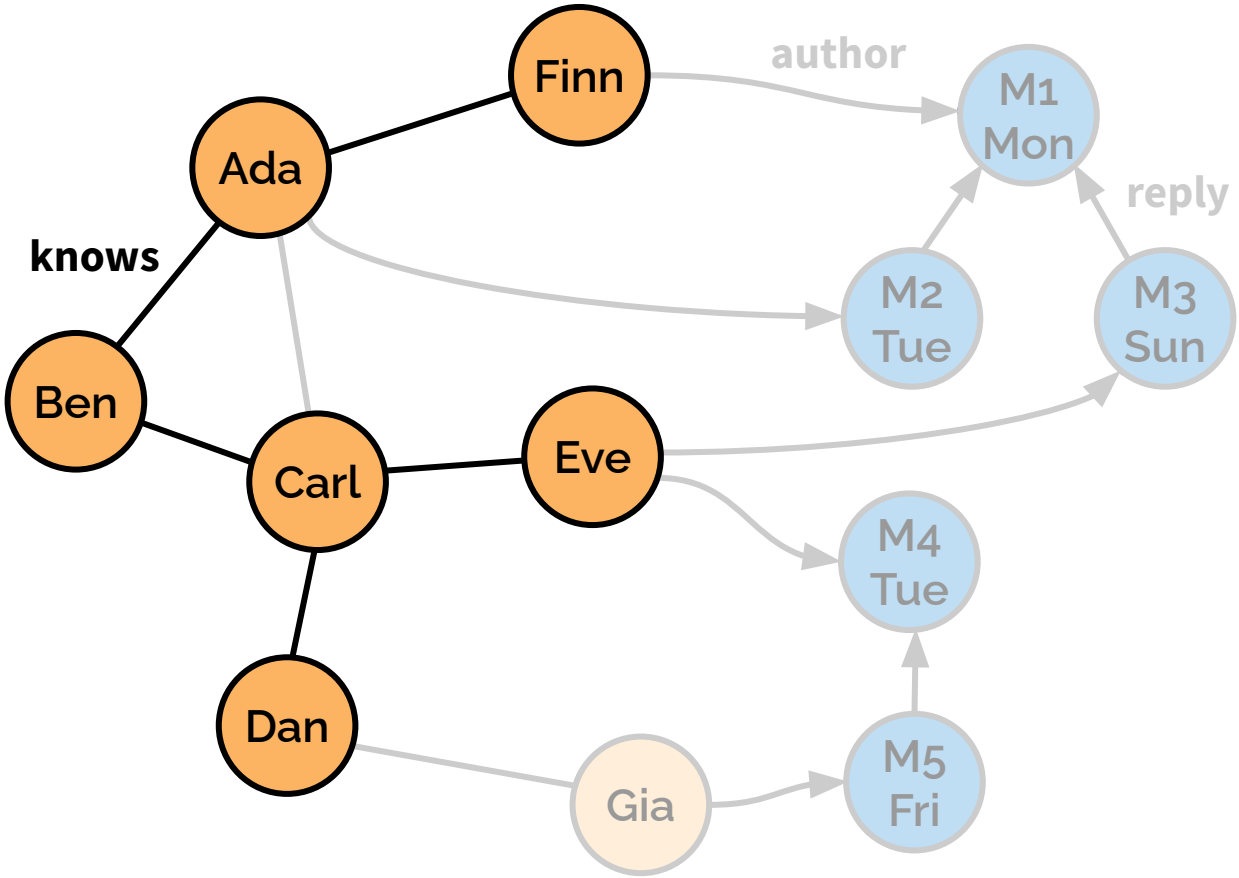
Data set | **Queries** | Updates



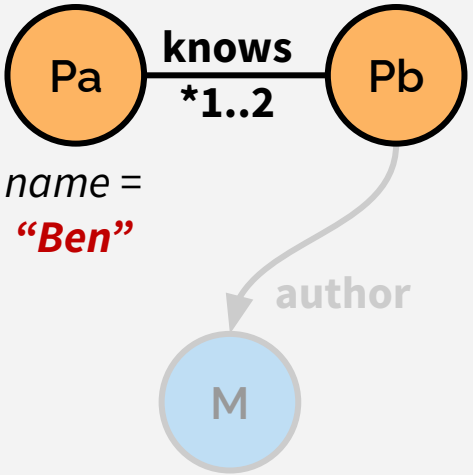
Q9(*"Ben"*, *"Sat"*)



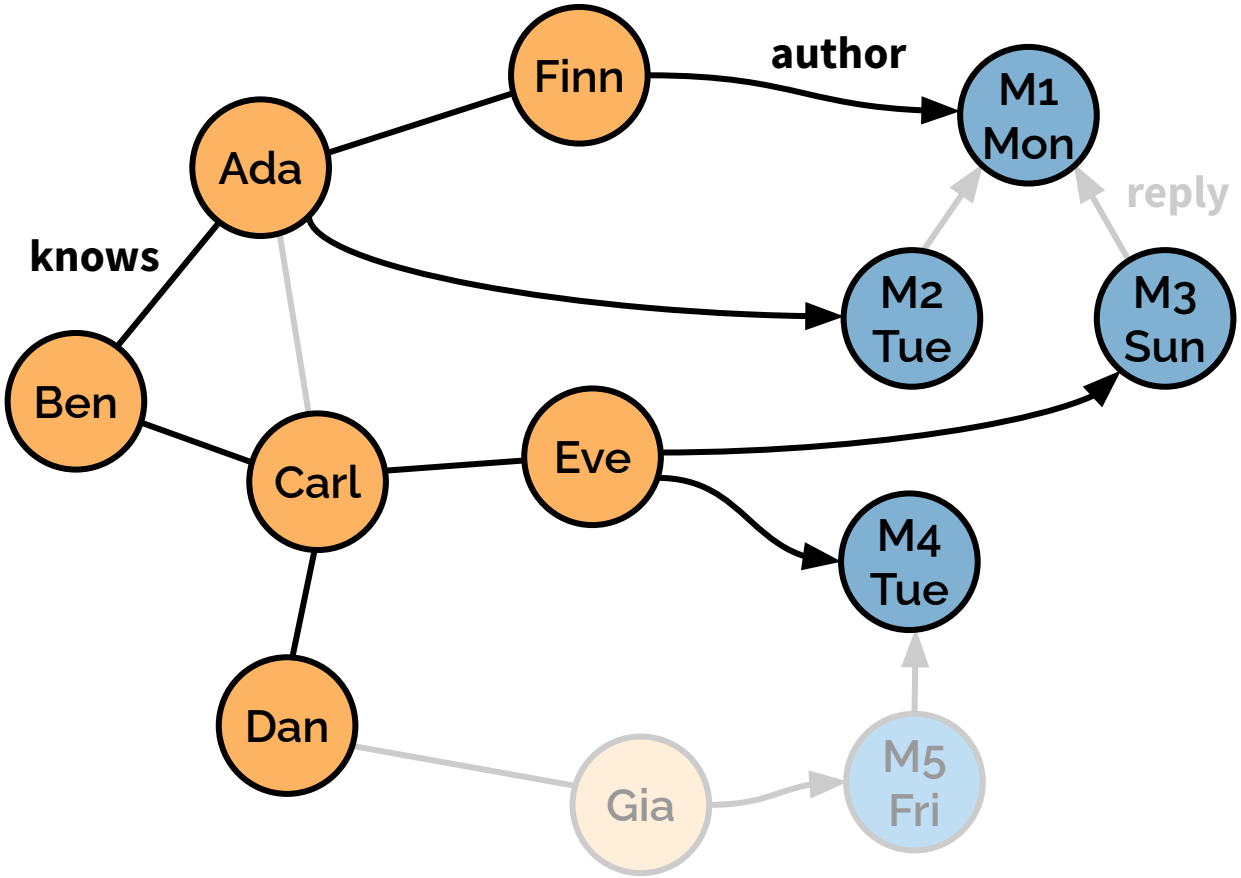
Data set | **Queries** | Updates



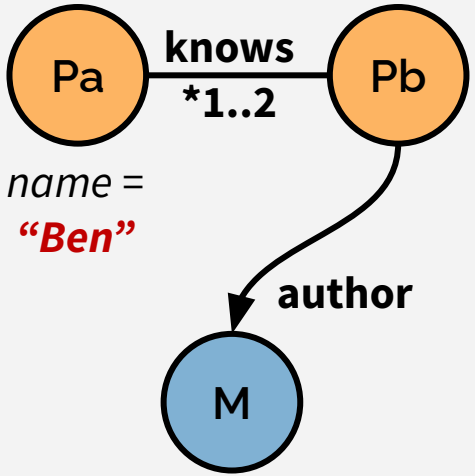
Q9(*"Ben"*, *"Sat"*)



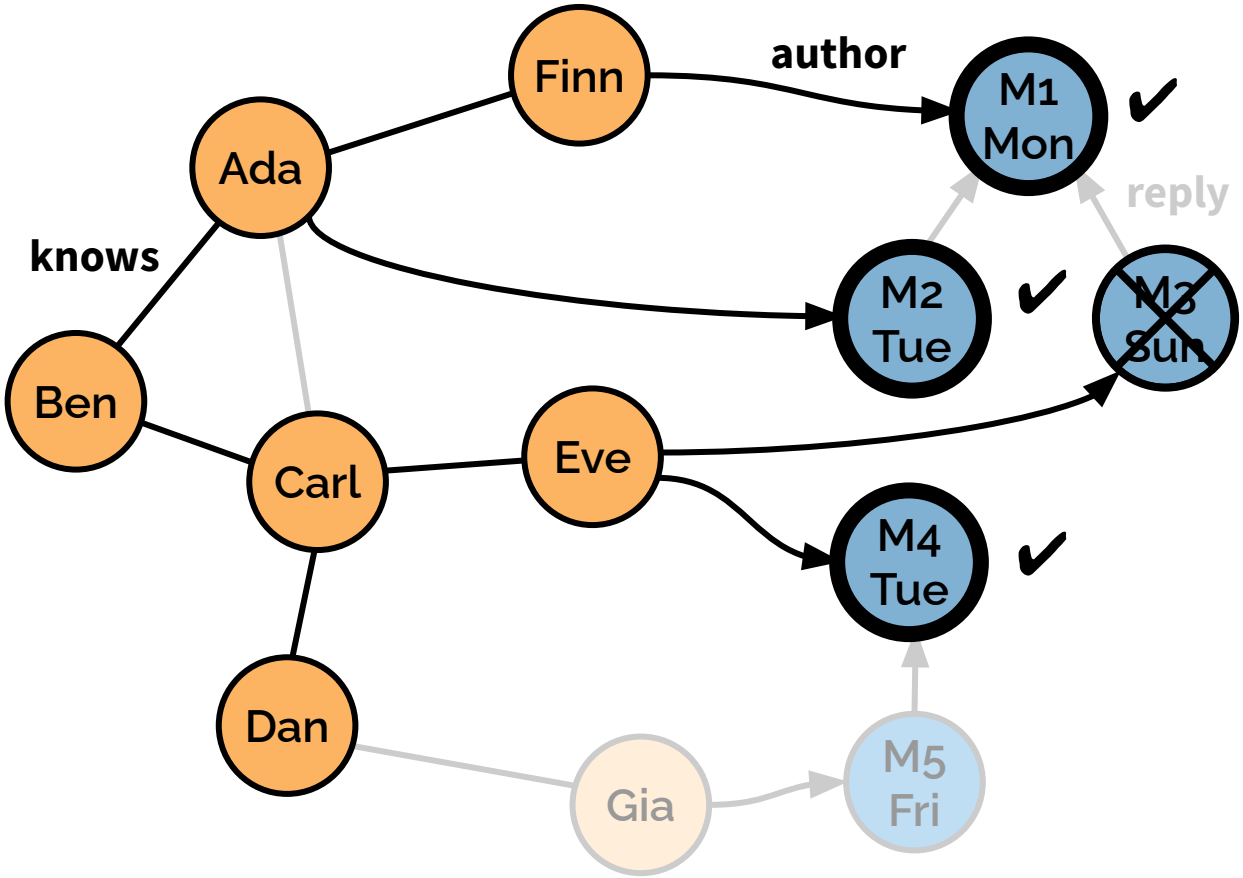
Data set | **Queries** | Updates



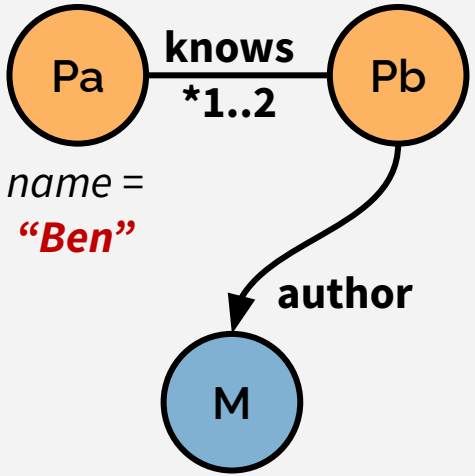
Q9(*"Ben"*, *"Sat"*)



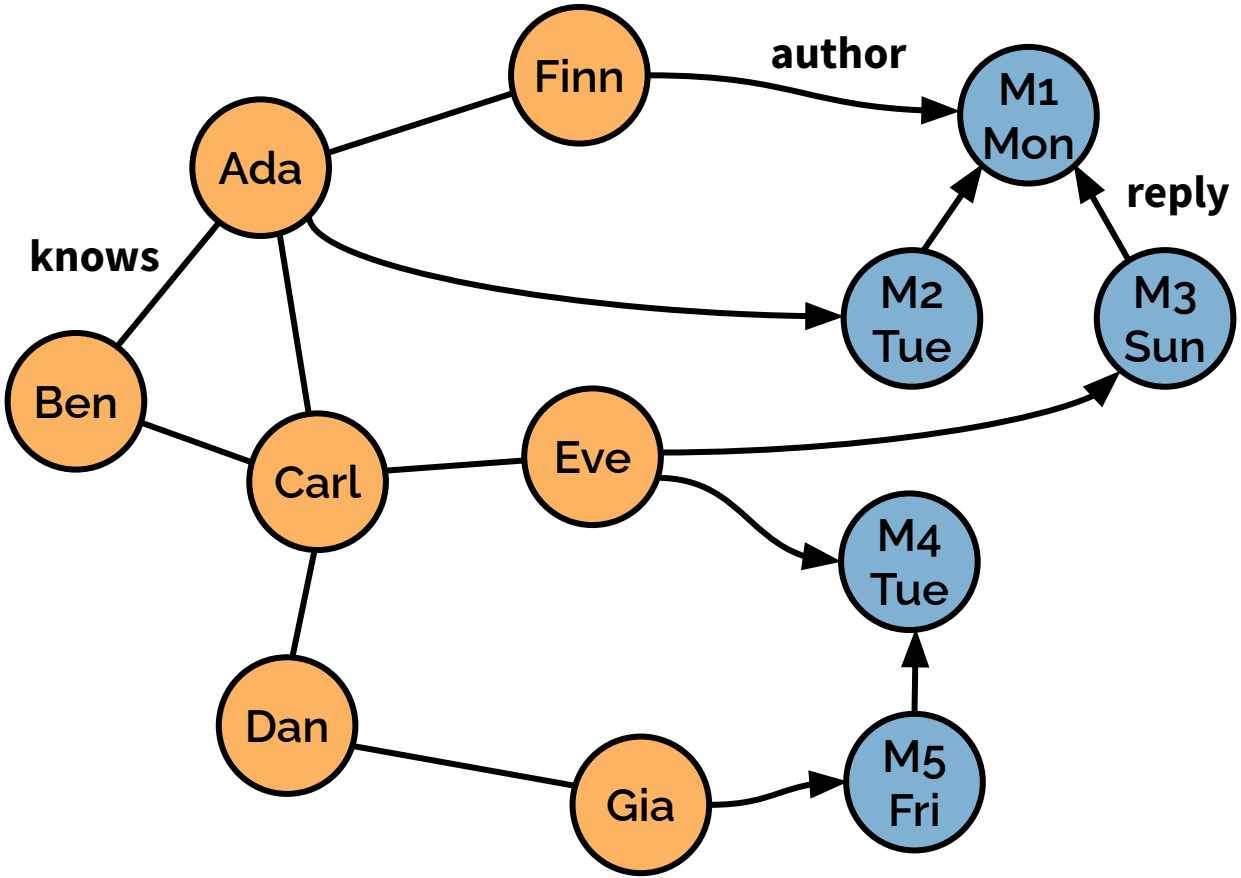
Data set | **Queries** | Updates



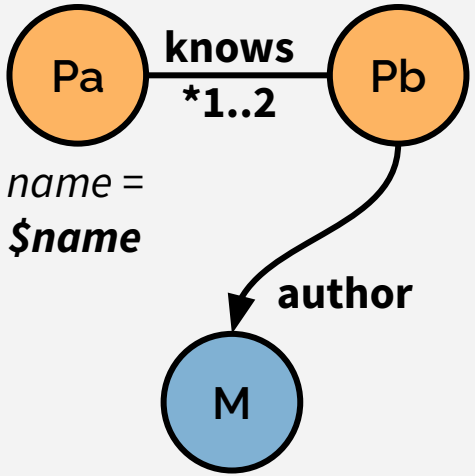
Q9(*"Ben"*, *"Sat"*)



Data set | **Queries** | Updates



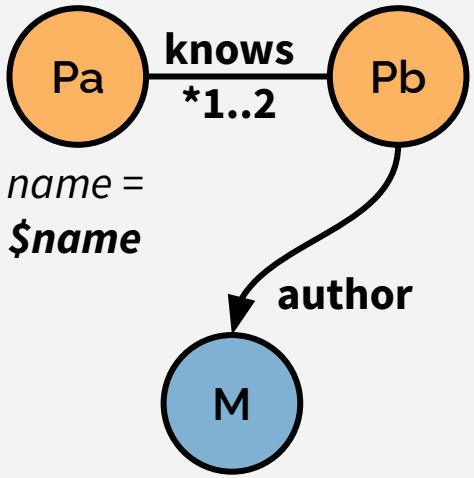
Q9(*$name*, *$day*)



SQL:1992

```
SELECT DISTINCT m.id
FROM (
 SELECT k.p2id AS id
 FROM person Pa,
      knows k
 WHERE Pa.name = $name
   AND Pa.id = k.p1id
 UNION
 SELECT k2.p2id AS id
 FROM person Pa,
      knows k1,
      knows k2
 WHERE Pa.name = $name
   AND Pa.id = k1.p1id
   AND k1.p2id = k2.p1id
   AND k1.p1id <> k2.p2id
 ) Pb,
 Message m
WHERE Pb.id = m.authorId
  AND m.creationDate < $day
```

Q9(*$name*, *$day*)



SQL/PGQ (SQL:2023)

```
SELECT id
FROM GRAPH_TABLE (socialNetwork
 MATCH ANY ACYCLIC
   (Pa:Person WHERE Pa.name = $name)
   -[:knows]-{1,2} (Pb:Person)
   -[:author]-> (m:Message)
 WHERE m.creationDate < $day
 COLUMNS (m.id))
```

GQL

```
MATCH ANY ACYCLIC
  (Pa:Person WHERE Pa.name = $name)
  -[:knows]-{1,2} (Pb:Person)
  -[:author]-> (m:Message)
WHERE m.creationDate < $day
RETURN DISTINCT m.id
```

Q13($src, $dst)

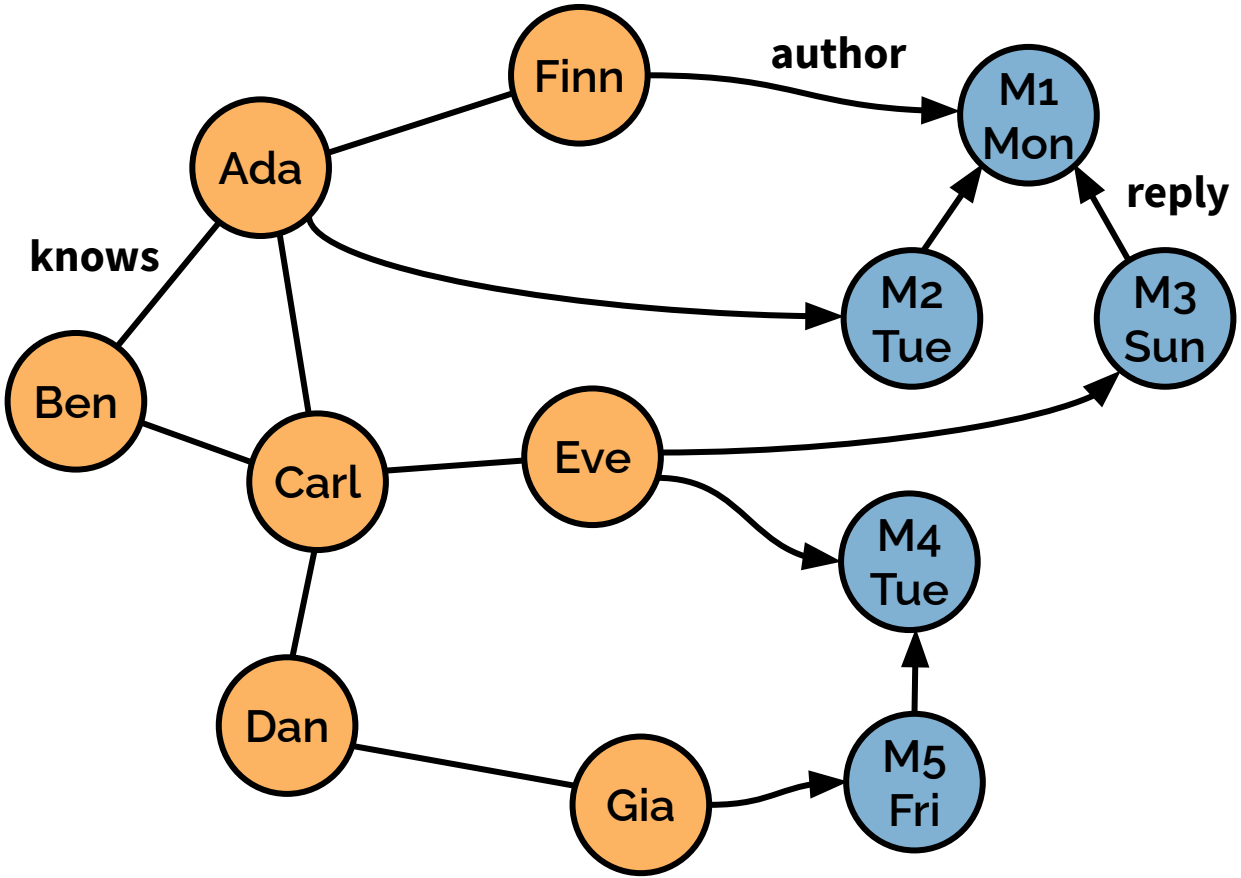
SQL/PGQ (SQL:2023)

```
SELECT length FROM GRAPH_TABLE (sn
 MATCH p = ANY SHORTEST
  (Pa:Person WHERE Pa.id = $src)-[:knows]-*
  (Pb:Person WHERE Pb.id = $dst)
COLUMNS (path_length(p) AS length))
```

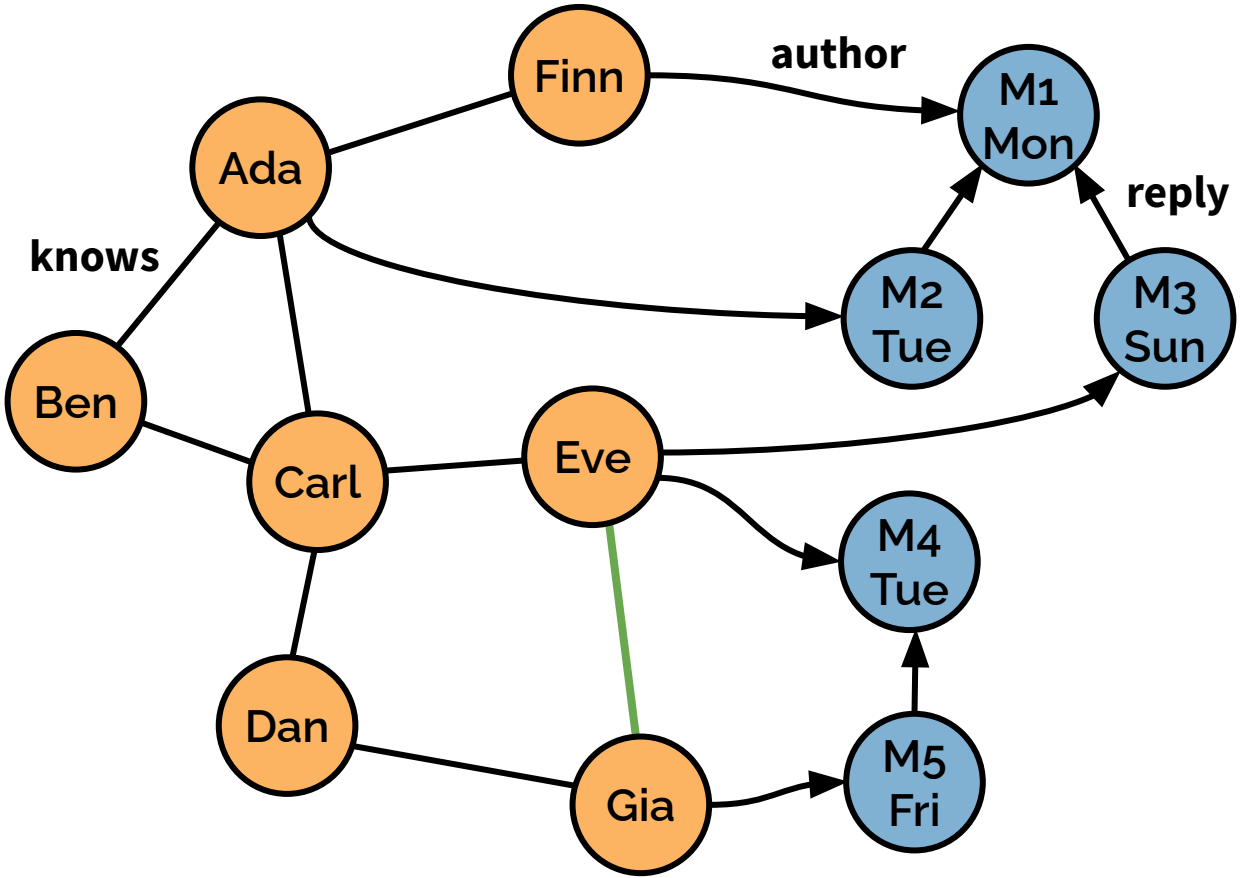
SQL:1999

```
WITH RECURSIVE ps(sp, ep, path, eR) AS (
  SELECT p1id AS sp, p2id AS ep, [p1id, p2id] AS path, (p2id = $dst) AS eR
    FROM knows WHERE sp = $src UNION ALL SELECT ps.sp AS sp, p2id AS ep,
    array_append(path, p2id) AS path, max(CASE WHEN p2id = $dst THEN 1 ELSE 0 END)
    OVER (ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS eR
    FROM ps JOIN knows ON ps.ep = p1id WHERE NOT EXISTS
    (SELECT 1 FROM ps pps WHERE list_contains(pps.path, p2id)) AND ps.eR = 0)
SELECT min(length(path)) AS length FROM ps WHERE ep = $dst
```

Data set | Queries | **Updates**



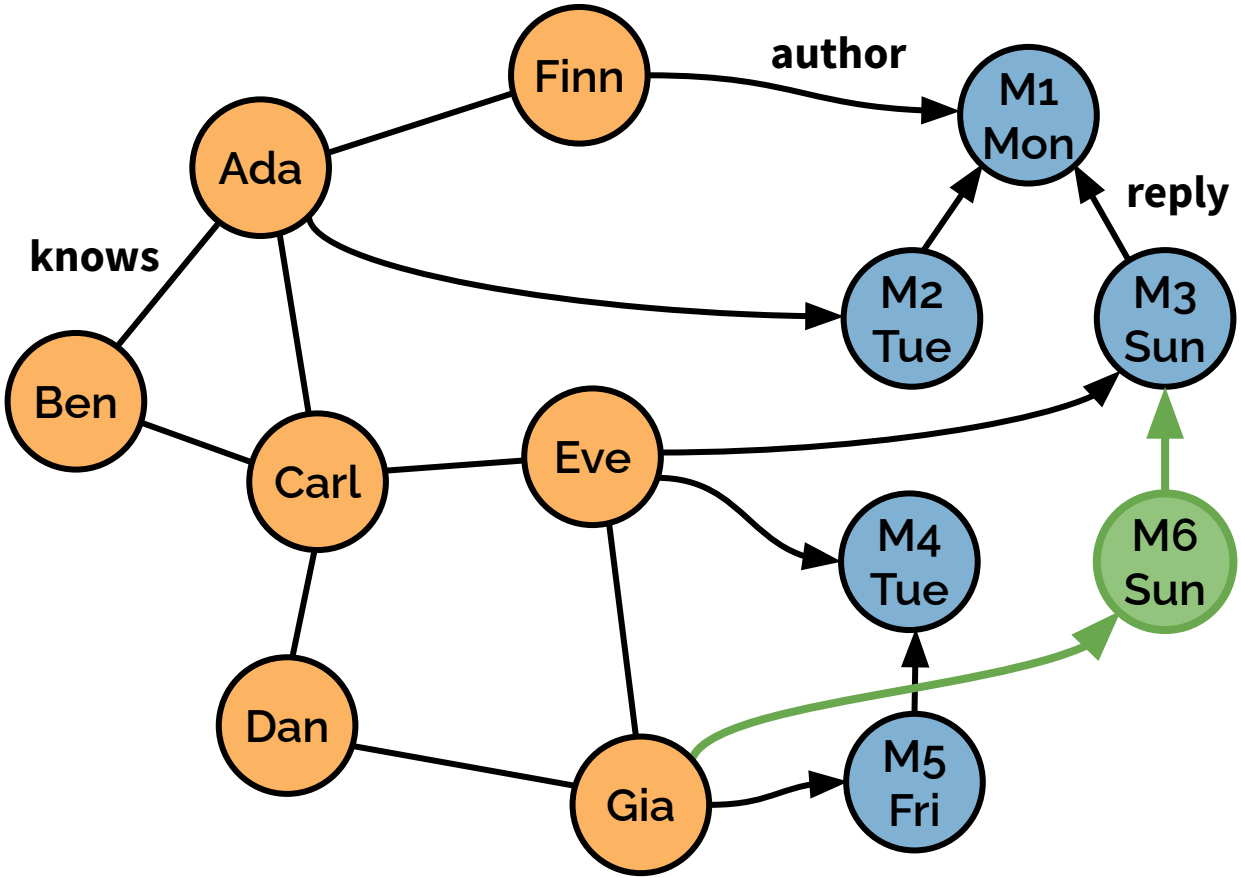
Data set | Queries | **Updates**



Updates

+ knows(“Eve”, “Gia”)

Data set | Queries | **Updates**

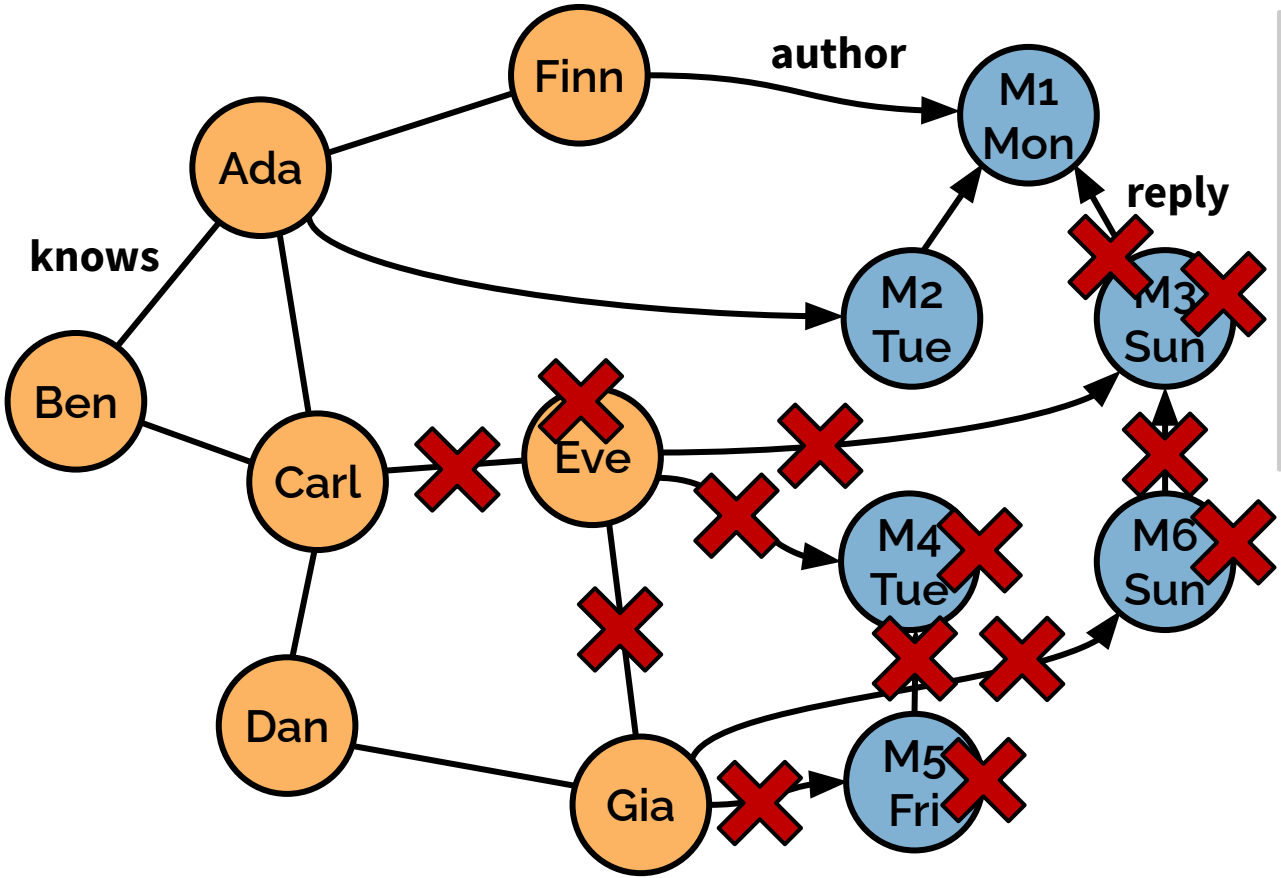


Updates

+ knows("Eve", "Gia")

+ Message("Gia", "M3")

Data set | Queries | **Updates**



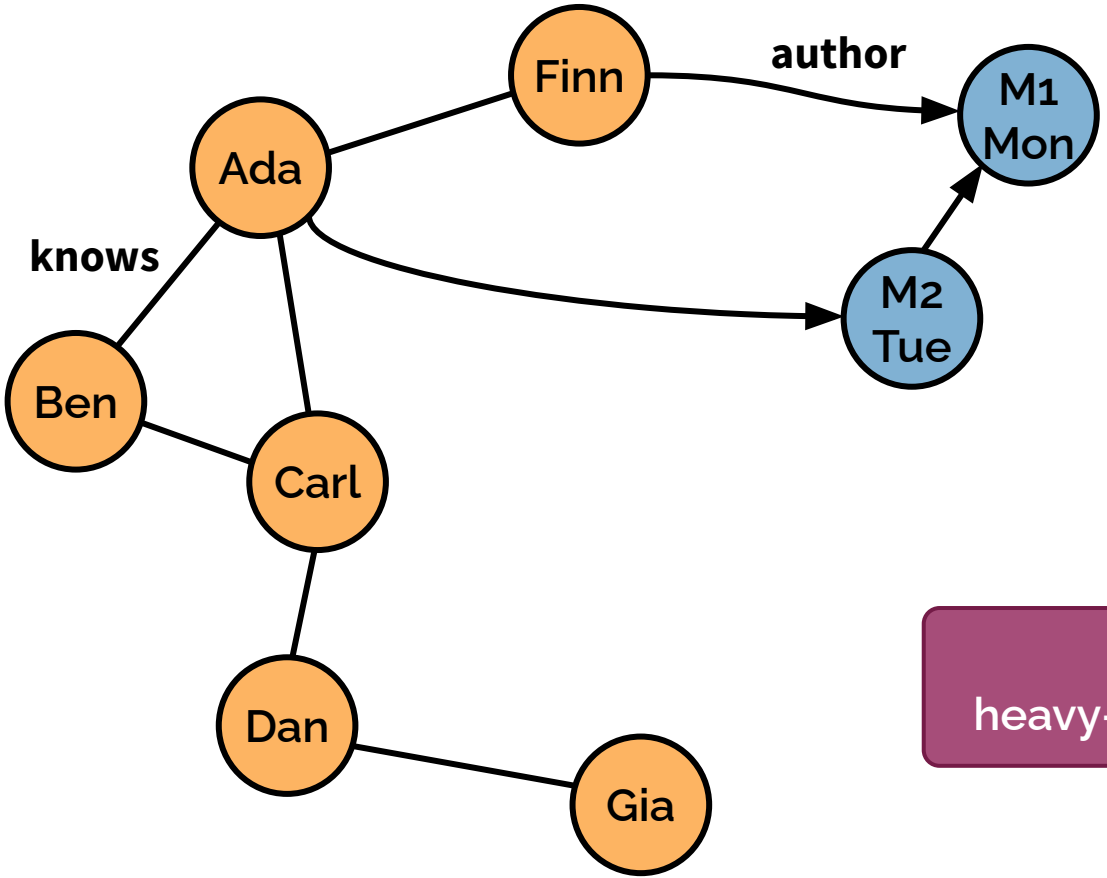
Updates

+ knows("Eve", "Gia")

+ Message("Gia", "M3")

– Person("Eve")

Data set | Queries | **Updates**



Updates

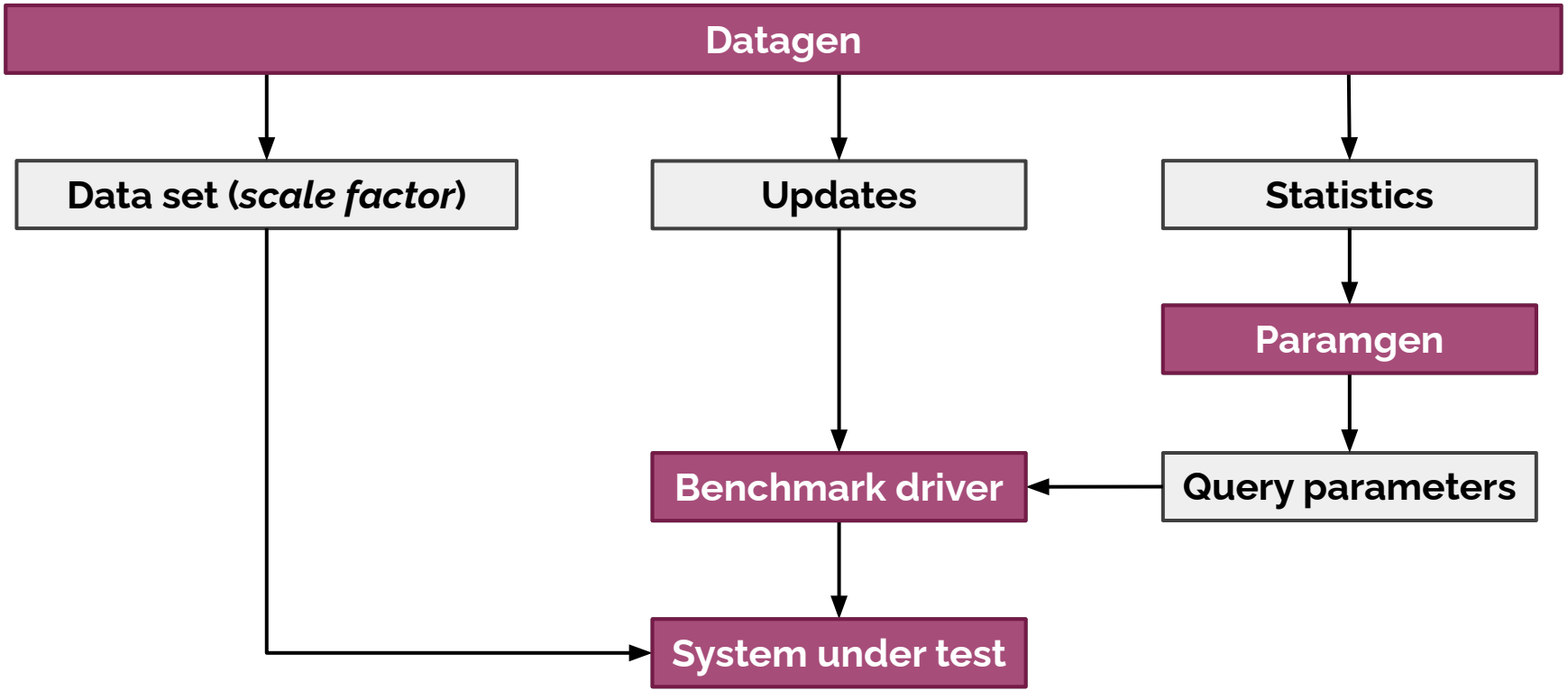
+ knows("Eve", "Gia")

+ Message("Gia", "M3")

– Person("Eve")

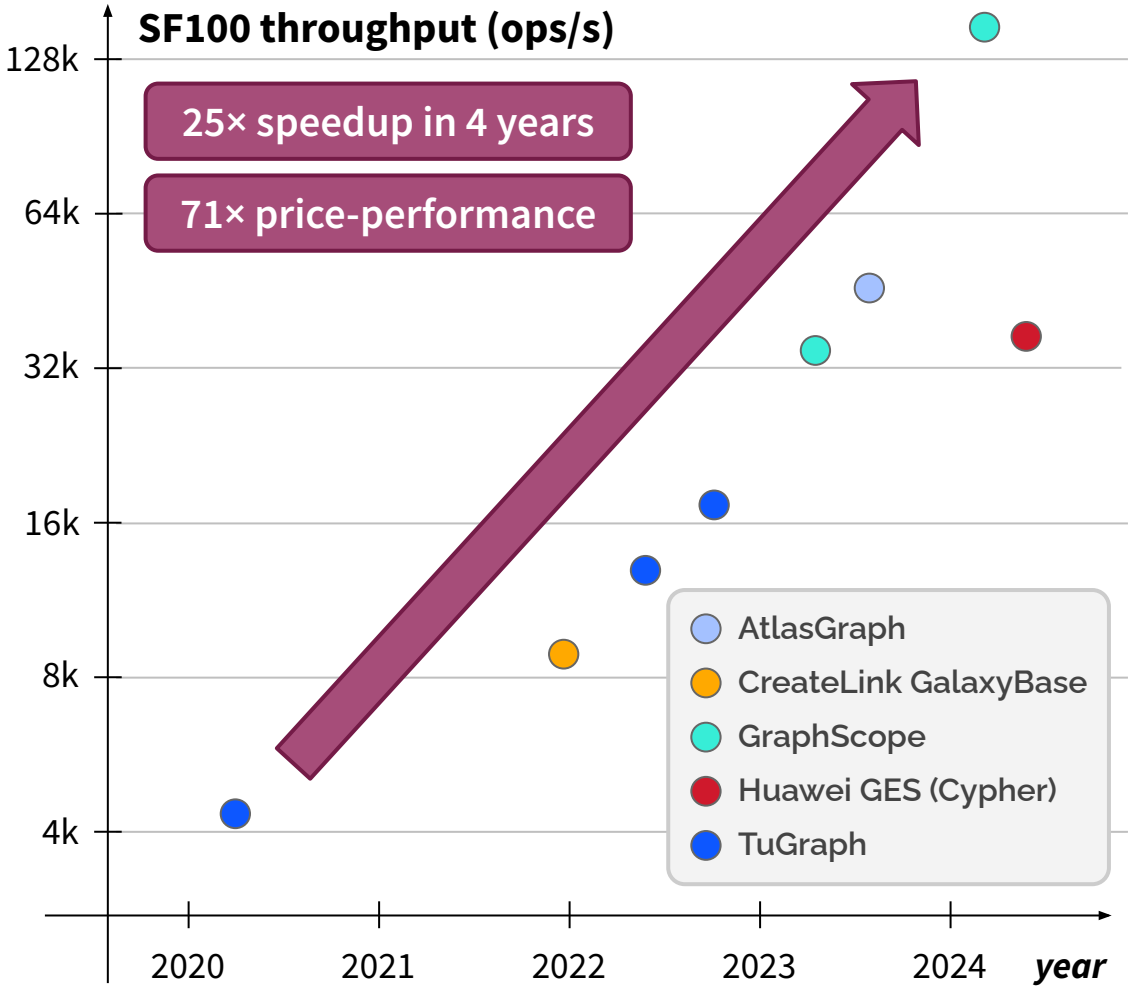
Deletes are heavy-hitting operations!

Benchmark workflow



SNB Interactive (2015)

- Transactional workload
- Target metric: throughput

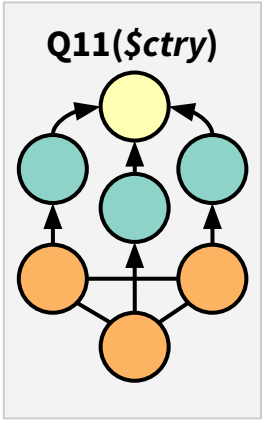


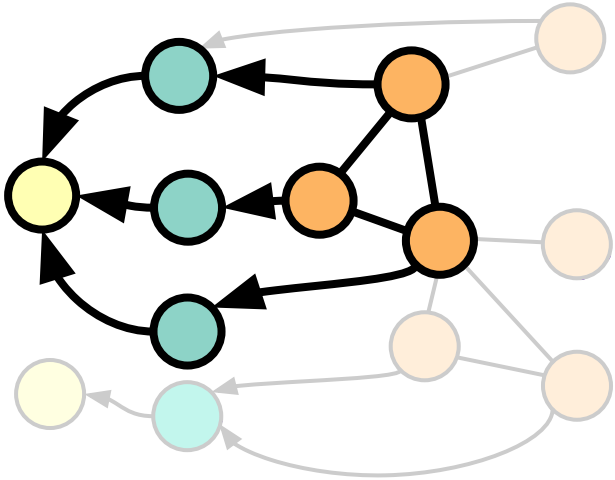
SNB Business Intelligence (2022)

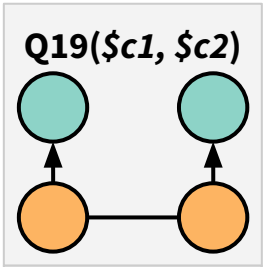
Analytical workload

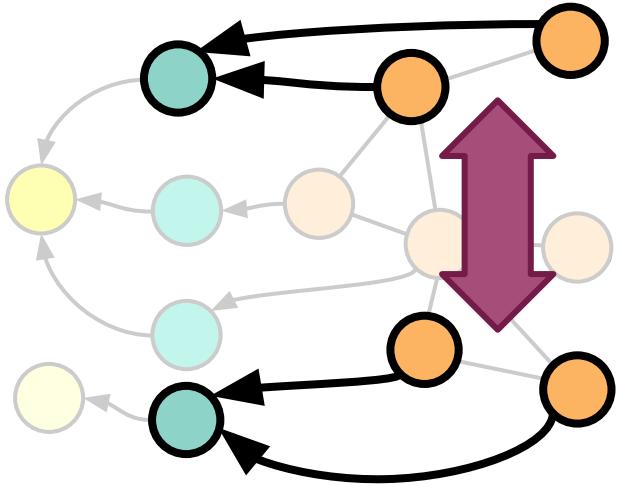
Metric 1: Power

Metric 2: Throughput









SNB Business Intelligence (2022)

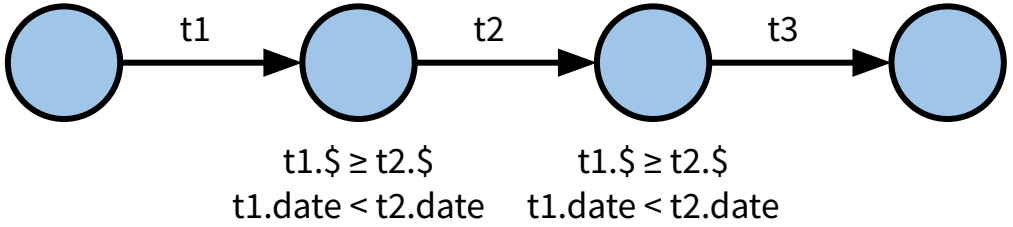
- Analytical workload
- Metric 1: Power
- Metric 2: Throughput

Audited results

	Scale factors
TigerGraph	100
	1,000 (×3)
	10,000
TuGraph™	30,000

Financial Benchmark (2023)

- Transactional workload
- Metric: Throughput
- Target: Distributed systems
- Relaxed consistency requirements



Financial Benchmark (2023)

- Transactional workload
- Metric: Throughput
- Target: Distributed systems
- Relaxed consistency requirements

Audited results

no audited results yet!

Using the benchmarks

Benchmark kit

Specification

Academic paper

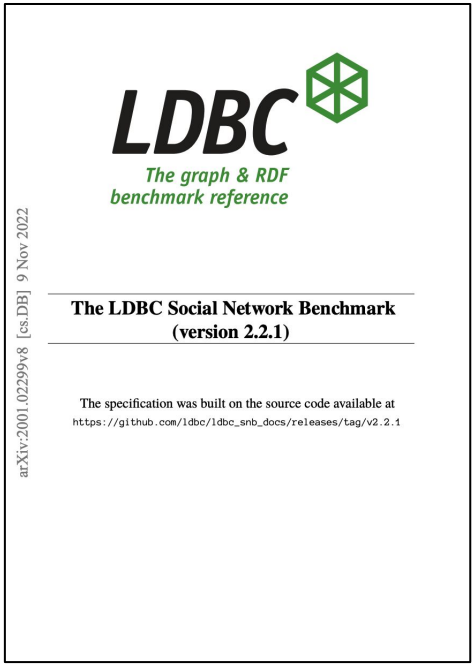
Data generator

Pre-generated data sets

Driver

2+ implementations

Guidelines



LDBC

The graph & RDF benchmark reference

The LDBC Social Network Benchmark (version 2.2.1)

The specification was built on the source code available at `https://github.com/ldbc/ldbc_snb_docs/releases/tag/v2.2.1`

The LDBC Social Network Benchmark: Business Intelligence Workload

Gábor Szárnyas, CWI, gabor.szarnyas@cwi.nl
Jack Waudby, Newcastle University, j.waudby2@ncl.ac.uk
Benjamin A. Steer, Pometry, ben.steer@pometry.com
Dávid Szakállas, Independent contributor, david.szakallas@gmail.com
Altan Birler, Technische Universität München, altan.birler@tum.de
Mingxi Wu, TigerGraph, mingxi.wu@tigergraph.com
Yuchen Zhang, TigerGraph, yuchen.zhang@tigergraph.com
Peter Boncz, CWI, boncz@cwi.nl

ABSTRACT

The Social Network Benchmark's Business Intelligence workload (SNB BI) is a comprehensive graph OLAP benchmark targeting analytical data systems capable of supporting graph workloads. This paper marks the finalization of almost a decade of research in academia and industry via the Linked Data Benchmark Council (LDBC). SNB BI advances the state-of-the art in synthetic and scalable analytical database benchmarks in many aspects. Its base is a sophisticated data generator, implemented on a scalable distributed infrastructure, that produces a social graph with small-world phenomena, whose value properties follow skewed and correlated distributions and where values correlate with structure. This is a temporal graph where all nodes and edges follow lifespan-based rules with temporal skew enabling realistic and consistent temporal inserts and (recursive) deletes. The query workload exploiting this skew and correlation is based on LDBC's "choke point"-driven design methodology and will entice technical and scientific improvements in future (graph) database systems. SNB BI includes the first adoption of "parameter curation" in an analytical benchmark, a technique that ensures stable runtimes of query variants across different parameter values. Two performance metrics characterize peak single-query performance (power) and sustained concurrent query throughput. To demonstrate the portability of the benchmark, we present experimental results on a relational and a graph DBMS. Note that these do not constitute an official LDBC Benchmark Result – only audited results can use this trademarked term.

PVLDB Reference Format:
Gábor Szárnyas, Jack Waudby, Benjamin A. Steer, Dávid Szakállas, Altan Birler, Mingxi Wu, Yuchen Zhang, and Peter Boncz. The LDBC Social Network Benchmark: Business Intelligence Workload. PVLDB, 16(4): 877 - 890, 2022.
doi:10.14778/3574245.3574270

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/ldbc/ldbc_snb_bi/releases/tag/v1.0.3.

Proceedings of the VLDB Endowment, Vol. 16, No. 4 ISSN 2150-8097.
doi:10.14778/3574245.3574270

Table 1: The SNB BI workload fills in the space between LDBC SNB Interactive and LDBC Graphalytics. It is a graph OLAP workload focusing on queries on a labelled attributed graph with temporal changes (inserts and deletes), targeting systems with domain-specific query languages. We denote the data models and features covered, and whether a language is *capable of implementing* and *allowed to implement* a given benchmark. *Notation:* ⊛: yes, ○: no, ◎: limited coverage.

	OLTP	OLAP	algorithms
LDBC benchmark	SNB Interactive	SNB BI	Graphalytics
labelled attributed graph	⊛	⊛	○
insert operations	⊛	⊛	○
delete operations	○	⊛	○
challenging joins	○	⊛	⊛
cheapest path finding	○	⊛	⊛
inter-query parallelism	required	optional	not allowed
query footprint	small	large	all data
SQL with recursion	⊛	⊛	○
GQL, SQL/PGQ, Cypher	⊛	⊛	○
GPQL	⊛	⊛	⊛
SPARQL-path extension	⊛	⊛	○
imperative API	⊛	○	⊛

1 INTRODUCTION

Analyzing the connection patterns in graphs is a steadily expanding use case in data analytics and is projected to still grow considerably in importance [57]. It is reflected in the increasing role of graph-shaped data as represented in data models such as (initially) RDF and increasingly *property graphs* [5]. While graph analytics is often associated with obviously graph-intensive application domains that manage data representing social networks, telecommunication networks, and enterprise knowledge graphs [60], graph challenges are also found in traditional relational data warehouses and modern data lakes, where implicit graphs lurk in the connection patterns formed between tables that refer to each other through joins along relationships, esp. along many-to-many relationships. Practitioners, data system builders, and researchers are increasingly focusing on graph analysis questions [56], performing tasks such as fraud detection, recommendation, historical analysis, and root-cause analysis.

The Linked Data Benchmark Council. To expedite the evolution of the modern graph data management stack, a group of industry and academic organizations founded the Linked Data Benchmark Council (LDBC) in 2012, originally as a European Union-funded project.

Auditing

Performed by certified auditors

Audited results are used in RFPs (Request for Proposals)

Benchmark setup	SF	Hardware	Performance	Performance (price-adjusted)
○ System: **GraphScope Flex 0.26.1** ○ Test sponsor: **Alibaba Cloud** ○ Date: **2024-05-14** ○ Queries implemented in: **C++** ○ System cost: **738,724 RMB** (102,128.22 USD)	100	Alibaba Cloud ecs.r8a.16xlarge 64×AMD EPYC 9T24 @ 3.7GHz vCPUs, 512GiB RAM	130,098.36 ops/s	1,273.873
	300	Alibaba Cloud ecs.r8a.16xlarge 64×AMD EPYC 9T24 @ 3.7GHz vCPUs, 512GiB RAM	131,263.87 ops/s	1,285.285

Total Cost of Ownership

We report the TCO based on the TPC Pricing Specification

3-year software license

3-year hardware / cloud serve

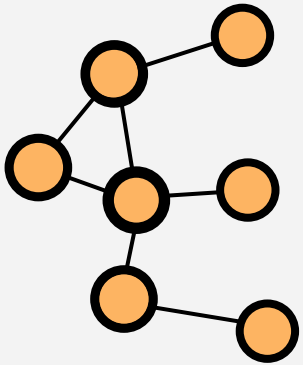
3-year maintenance (enterprise-grade support):

- *7 days/week, 24 hours/day coverage*
- *“the response time for problem recognition* ***must not exceed 4 hours****”*

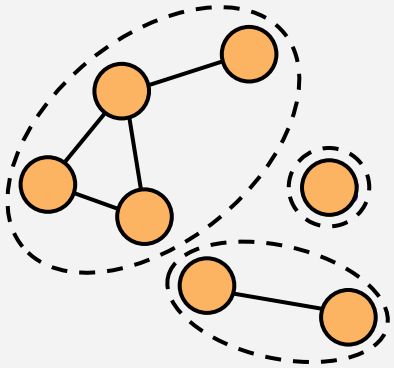
Read-only workloads

Graphalytics (2016)

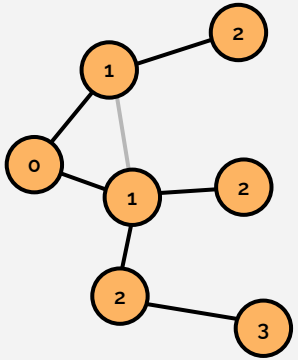
- Graph algorithms
- Macrobenchmark
- Unlabelled, unattributed graphs
- Metric: Processing time



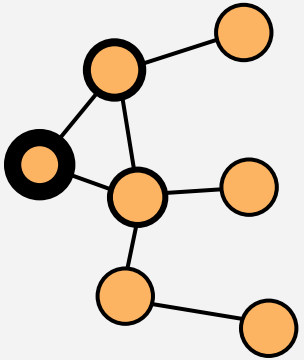
PageRank



Weakly CC



BFS



Local clustering

Labelled Subgraph Query Benchmark (2021)

Graph pattern matching

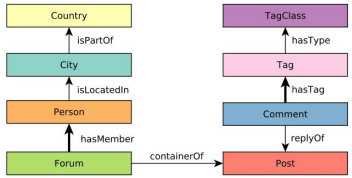
Metric: Total runtime

Focus: Research

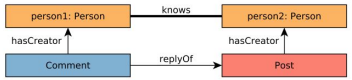
Labelled, unattributed graph

Audited results

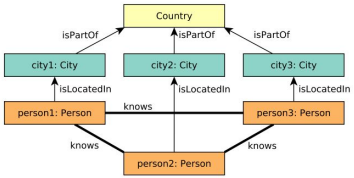
n/a



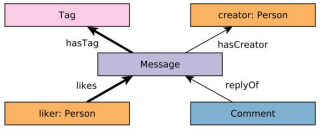
(a) Q1.



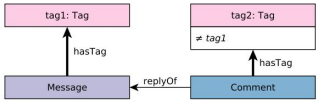
(b) Q2.



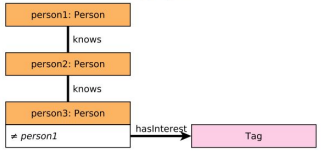
(c) Q3.



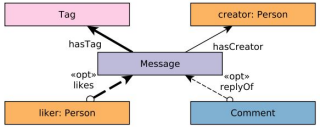
(d) Q4.



(e) Q5.

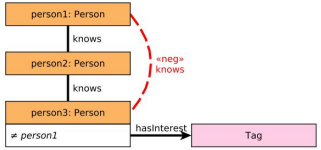


(f) Q6.



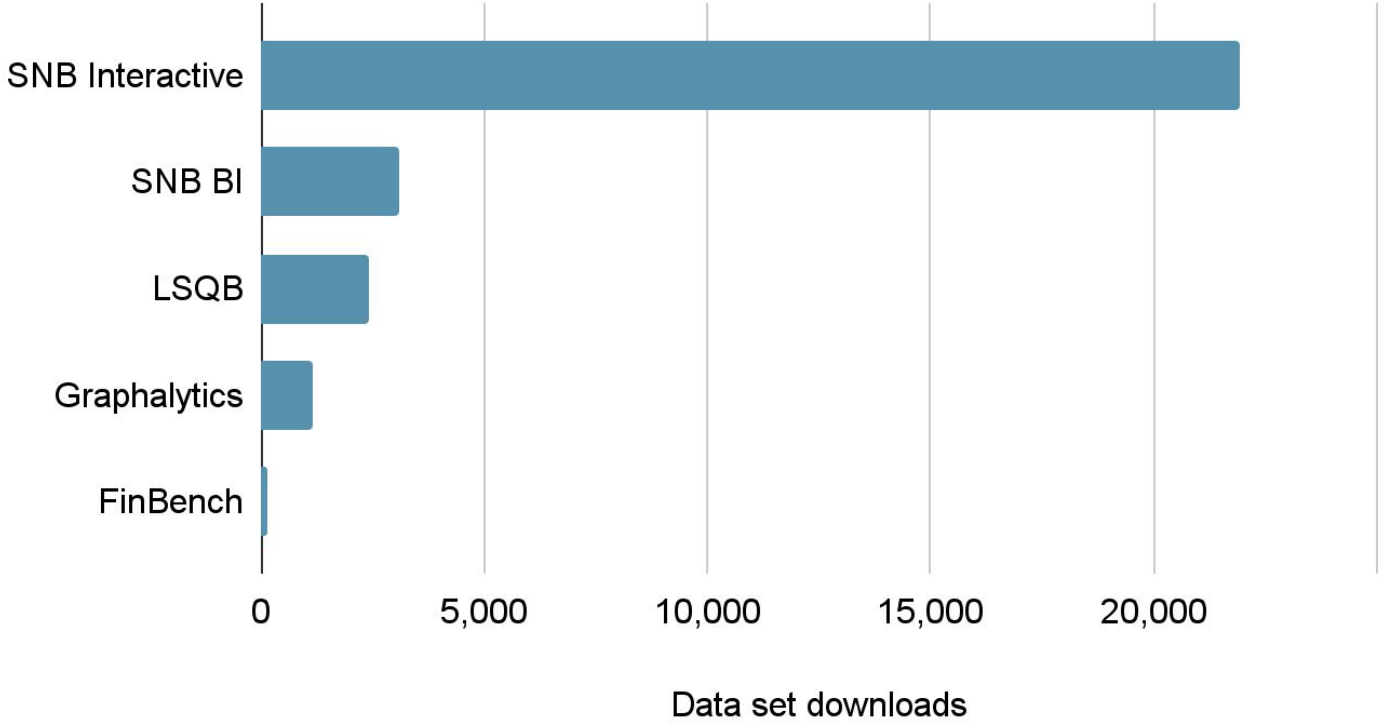
(g) Q7.

(h) Q8.



(i) Q9.

Usage statistics

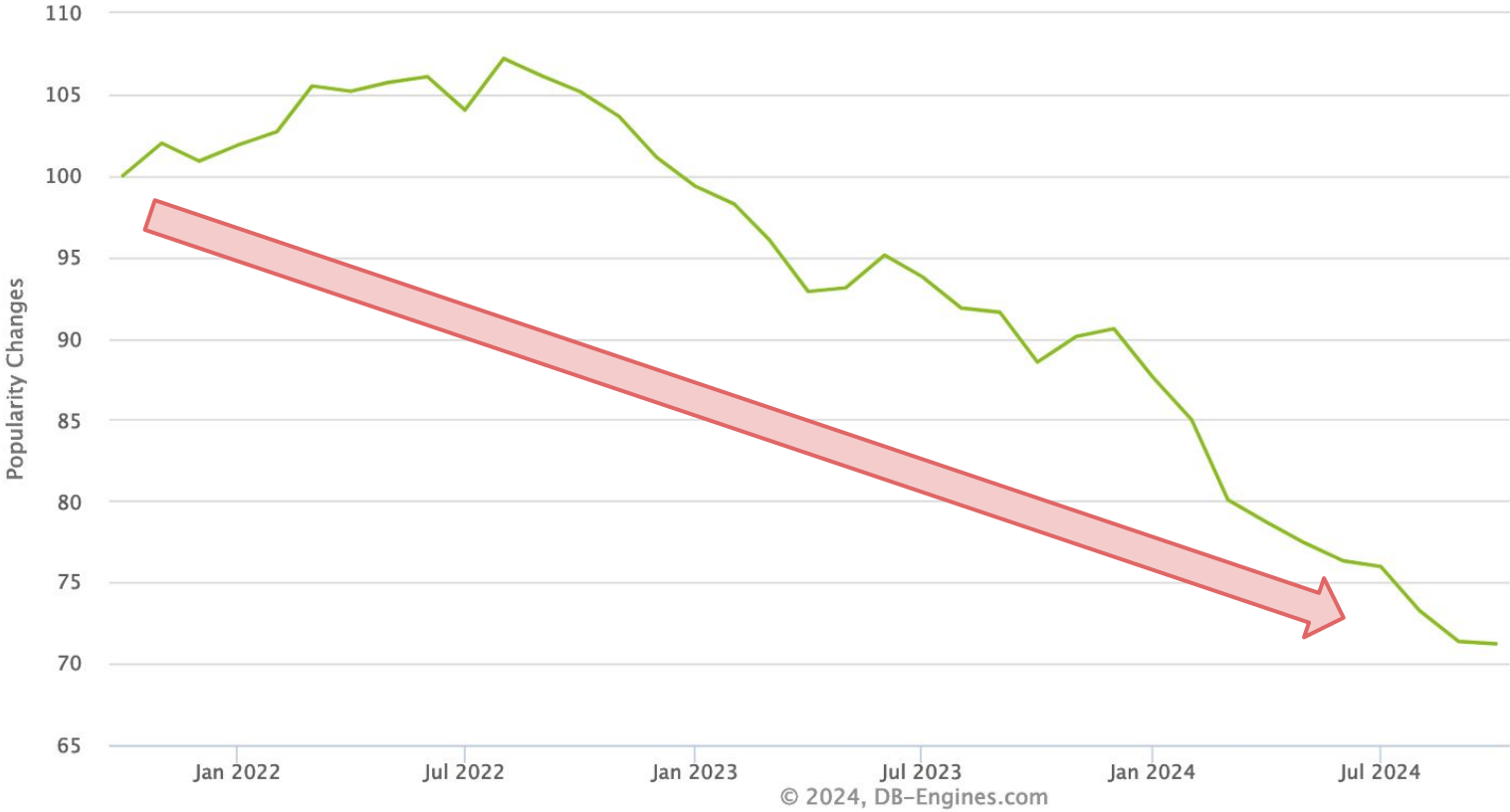


Comparison with TPC benchmarks

macro / application-level benchmarks	“scale factors”: SF30 = 30GiB CSV	flexible hardware and software setup
auditor training, exam, and certification	competing on metrics, e.g. throughput	benchmark approval and renewal
only members can commission audits	reports are written by auditors	no standard query language required

Challenges
in the graph database space

DB Engines Ranking for graph: ¼ drop in 3 years



Areas for LDBC to improve in

Covering important recent technologies

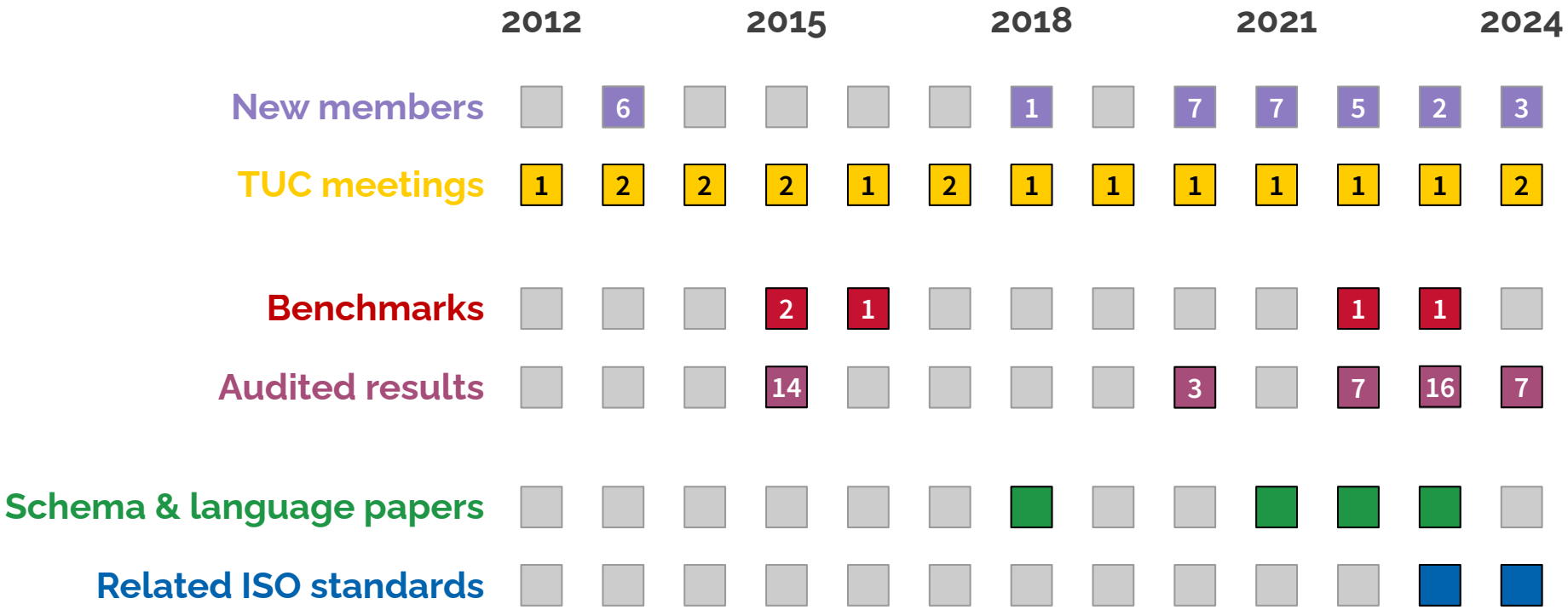
Cloud infrastructure and cloud-native systems

- serverless setups
- take elasticity into account for pricing

ML workloads

- graph neural networks
- knowledge graphs
- vector databases

LDBC: Summary of 12 years


2012
2015
2018
2021
2024
New members
6
1
7
7
5
2
3
TUC meetings
1
2
2
2
1
2
1
1
1
1
1
1
2
Benchmarks
2
1
1
1
Audited results
14
3
7
16
7
Schema & language papers
Related ISO standards


LDBC
The graph & RDF
benchmark reference