THE WORLD OF GRAPH DATABASES
- FROM AN INDUSTRY PERSPECTIVE

Yuanyuan Tian
Gray Systems Lab, Microsoft
A bit about myself related to graphs

• Long standing interest in graphs since 2003
  • Two books on graphs
  • 19/50+ publications on graphs (2700+ citations)

• Current: Principal Scientist Manager @ GSL
  • Working on graph projects with Azure Data and Liquid Team @ LinkedIn

• Past: Principal Research Staff Member @ IBM Research
  • The tech lead for IBM Db2 Graph product
• The global graph analytics market in 2022 was valued at **$1.14B**

• Global market for graph databases will grow at projected CAGR of **34.8%** during forecasted period (2023-2028) and reach **$6.9B** by 2028.
Customer Use Cases

- Finance
- Insurance
- Healthcare
- Security
- Retail
- Energy
- Power
- Manufacturing
- Supply chains
- Transportation
Graph Workloads

**Graph Queries (Graph OLTP)**
low-latency graph traversal and pattern matching

- e.g. neighbors of a vertex, shortest path between two vertices

**Graph Algorithms (Graph OLAP)**
iterative, long running, graph processing

- e.g. Pagerank, community detection

**New Trend: Graph ML/AI**
GNN

- e.g. node embedding, graph embedding

Vector
Graph Models

Resource Description Framework (RDF)
• Directed, edge-labeled graph (subject-predicate-object triple)
• Application
  • Knowledge representation & inference
  • Semantic Web

Property Graphs (PG)
• Directed, vertex-labeled, and edge-labeled graph with properties on each vertex/edge
• Application
  • Graph traversals/pattern matching
  • Path/graph analytics

Most graph databases support PG model

Patient 1
  hasID 198076
  hasName Alice Brown

03/24/2020
  happensOn Diagnosis 1

Diagnosis 1
  diagnosedWith Type 2 Diabetes
  hasID 64572326
  hasName

Type 2 Diabetes
  hasID 64572345
  hasName

Diabetes
  hasName

Disease 2
  isa

Properties: ID = 198076
  name = “Alice Brown”

Properties: ID = 64572326
  name = “Type 2 diabetes”

Properties: ID = 64572345
  name = “Diabetes”
Graph languages/Interfaces

Graph OLTP

- SPARQL for RDF graphs
- Chaos for PG graphs
  - Gremlin (imperative)
    - Supported by ~30 graph vendors
  - openCypher (declarative)
    - Supported by ~10 graph vendors
  - ISO Standard efforts: GQL and SQL/PGQ (declarative)

Need a few years to settle down!

Choose a standard/popular language and avoid inventing new ones!

Graph OLAP

- API/DSL + built-in algorithms
  - Most support Pregel-like API

https://www.gqlstandards.org/existing-languages
Graph Technology Landscape 2022

<table>
<thead>
<tr>
<th>Enterprise Cloud Companies</th>
<th>Deployment</th>
<th>Graph Model</th>
<th>Query Language</th>
<th>Visualization tools</th>
<th>Transaction</th>
<th>Graph OLAP</th>
<th>Scale-Out</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amazon Neptune</strong></td>
<td>AWS</td>
<td>PG, RDF</td>
<td>GSQL, SPARQL</td>
<td>Neptune Workbench</td>
<td>ACID</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Microsoft SQL Graph</strong></td>
<td>On-prem / Azure</td>
<td>PG</td>
<td>SQL Extension</td>
<td>Power BI plugin, 3rd party tools</td>
<td>ACID</td>
<td>Python/R scripts via Machine Learning Services</td>
<td>Yes* (Read-Only Queries)</td>
</tr>
<tr>
<td><strong>Microsoft Cosmos DB Graph</strong></td>
<td>Azure</td>
<td>PG</td>
<td>Gremlin</td>
<td>Azure Portal, 3rd party tools</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Oracle Spatial and Graph</strong></td>
<td>On-prem / OCI AWS, Azure, GCP</td>
<td>PG, RDF</td>
<td>PGQL, SPARQL</td>
<td>Graph Studio</td>
<td>ACID</td>
<td>Green Marl DSL, 50+ built-in algorithms (including Graph ML)</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>IBM Db2 Graph</strong></td>
<td>On-prem / CP4D</td>
<td>PG</td>
<td>Gremlin</td>
<td>Graph UI</td>
<td>ACID</td>
<td>-</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Graph Only Companies**

<table>
<thead>
<tr>
<th>Deployment</th>
<th>Graph Model</th>
<th>Query Language</th>
<th>Visualization tools</th>
<th>Transaction</th>
<th>Scale-Out</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TigerGraph</strong></td>
<td>On-prem / AWS, Azure, GCP</td>
<td>PG</td>
<td>GSQL</td>
<td>Graph Studio</td>
<td>ACID</td>
</tr>
<tr>
<td><strong>Neo4J</strong></td>
<td>On-prem / AWS, Azure, GCP</td>
<td>PG</td>
<td>Cypher</td>
<td>Studio</td>
<td>Non-repeatable reads may occur</td>
</tr>
<tr>
<td><strong>DataStax Enterprise Graph</strong></td>
<td>On-prem / AWS, Azure, GCP</td>
<td>PG</td>
<td>Gremlin</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Databricks GraphX &amp; GraphFrames</strong></td>
<td>On-prem / AWS, Azure, GCP</td>
<td>PG</td>
<td>Motif Finding DSL</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Competitive Landscape**
Graph Solution Space

Native Graph DB
- Everything from scratch
- Pros: performance
- Cons: high engineering cost

Hybrid Graph DB
- Graph engine + existing backend store
- Pros: faster development, leverage backend store
- Cons: performance
Graph Solution Space

Graph-only DB
• Only support graph workload (con)

Converged DB
• Support poly query languages/APIs on the shared data (pro)

Native Graph DB
Neo4j
TigerGraph

Converged DB (Multi Model)
IBM Db2 Graph
Oracle Spatial & Graph
DataStax Graph
Microsoft SQL Graph
Microsoft Cosmos DB Graph

AWS Neptune
Hybrid Graph DB
Advantage of Converged DB solution

Poly languages/APIs on shared data
• View the data in the way that is needed!
• No data transfer or transformation cost
• *If graph queries on original data* (no schema change, no secondary copy)
  • No disturbance of existing applications
  • Transaction updates are visible to graph analysis in real time

Leverage of existing backend data store
• Transaction support
• Access control
• Compliance to audits and regulations
• Temporal support
• Scalability
• HA & DR
Graph Benchmarks

- Graph500 Benchmark
- HPC Scalable Graph Analysis Benchmark
- LinkBench
- Open Graph Benchmark
- **LDBC Benchmarks (most comprehensive)**
  - LDBC-SNB (used by TigerGraph and Neo4j)
  - LDBC Graphalytics
  - LDBC SPB
  - FinBench

*All future performance studies should adopt LDBC benchmarks!*
Opportunities and Directions

Growing market for graph databases (CAGR 34.8%, $6.9B by 2027)

Graph-only vendors are currently leading
• Strength: performance and algorithm support
• Weakness: Data import/export is a bottleneck for end-to-end scenarios

Major cloud vendors are investing in graph space
• Advantage: they own the whole stack, including the source of truth
Recommendation for Researchers (more practical impact)

• Use widely-adopted graph models, languages, and benchmarks

• Practical challenges that industry faces:
  • Multi-tenancy and access control
  • Security and compliance
  • End-to-end pipelines with mixed graph and non-graph workloads
  • Dynamic graphs
Questions & Suggestions?