Graphalytics: A Big Data Benchmark for Graph-Processing Platforms

Mihai Capotă, Tim Hegeman, Alexandru Iosup, Arnau Prat-Pérez, Orri Erling, Peter Boncz

Delft University of Technology
Universitat Politècnica de Catalunya
OpenLink Software
CWI
The data deluge: large-scale graphs

Tens of billions of edges
The data deluge: large-scale graphs

Linkedin

Amazon

Spotify

Twitter

Steam

Graph processing

Tens of billions of edges
Platform diversity

- Hadoop
- Neo4j
- Apache Giraph
- GraphX
- Oracle Labs PGX
- GraphLab
Yet another benchmark?
Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms
Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms
- Graph500
  - BFS
  - Kroneker graph
Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms
- Graph500
  - BFS
  - Kroneker graph
- Several academic studies
  - Specific to graph or RDF databases
  - Ad hoc setup, difficult to extend
P-A-D triangle

Algorithm

P
Platform

A

D
Dataset
P-A-D triangle

Performance is enabled, portability is disabled
P-A-D triangle

Performance is enabled, portability is disabled

Algorithms for different data types and graphs

Algorithm

Platform

Dataset
P-A-D triangle

Performance is enabled, portability is disabled

Algorithms for different data types and graphs

No systematic findings yet
P-A-D triangle

Performance is enabled, portability is disabled

Algorithm

Deployment

No systematic findings yet

Platform

Algorithms for different data types and graphs

Dataset
Graphalytics
Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
Graphalytics

• The first comprehensive benchmark for big data graph-processing platforms
  – Advanced benchmark harness
  – Choke-point analysis
Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
  - Choke-point analysis
  - Realistic graph generator
Graphalytics

• The first comprehensive benchmark for big data graph-processing platforms
  – Advanced benchmark harness
  – Choke-point analysis
  – Realistic graph generator
Graphalytics

• The first comprehensive benchmark for big data graph-processing platforms
  – Advanced benchmark harness
  – Choke-point analysis
  – Realistic graph generator

• Co-sponsored by Oracle
Choke points
Choke points

• Technological challenges for platforms
Choke points

• Technological challenges for platforms
• Identified by experts
Choke points

- Technological challenges for platforms
- Identified by experts
- Real-world scenarios are enhanced to stress choke points
  - Prevent tunnel vision
  - Advance state of the art
Choke points

- Technological challenges for platforms
- Identified by experts
- Real-world scenarios are enhanced to stress choke points
  - Prevent tunnel vision
  - Advance state of the art

Erling et al., The LDBC Social Network Benchmark: Interactive Workload, SIGMOD 2015
Examples of choke points
Examples of choke points

• Network utilization
Examples of choke points

- Network utilization
- Memory footprint
Examples of choke points

- Network utilization
- Memory footprint
- Access locality
Examples of choke points

- Network utilization
- Memory footprint
- Access locality
- Skewed execution
Examples of choke points

- Network utilization
- Memory footprint
- Access locality
- Skewed execution
- CPU?
Examples of choke points

● Network utilization
● Memory footprint
● Access locality
● Skewed execution
● CPU?

Ousterhout et al., “Making Sense of Performance in Data Analytics Frameworks”, NSDI 2015
Examples of choke points

- Network utilization
- Memory footprint
- Access locality
- Skewed execution
- CPU?
- Others?

Ousterhout et al., “Making Sense of Performance in Data Analytics Frameworks”, NSDI 2015
Realistic graph generator
Realistic graph generator

• LDBC Datagen
  - Synthetic social network similar to Facebook
Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook

- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented
Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook
- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented
  - Other structural characteristics
    - Clustering coefficient
    - Assortativity
Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook

- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented
  - Other structural characteristics
    - Clustering coefficient
    - Assortativity

Erling et al., The LDBC Social Network Benchmark: Interactive Workload, SIGMOD 2015
Advanced benchmark harness

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

Results

System Monitor

Graph processing platform

Dataset

Generator

39
Advanced benchmark harness

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

Results

System Monitor

Graph processing platform
Advanced benchmark harness
Advanced benchmark harness

- **User**
- **Datasets**
  - **Dataset Generator**
  - **Benchmark Core**
    - **Platform-specific algorithm implementation**
  - **Configuration**
  - **Report Generator**
  - **Output Validator**
  - **System Monitor**
  - **Graph processing platform**
  - **Results**
Advanced benchmark harness

- **User**
- **Datasets**
  - Dataset Generator
  - Platform-specific algorithm implementation
- **Benchmark Core**
- **Configuration**
- **Report Generator**
- **Output Validator**
- **System Monitor**
- **Graph processing platform**
- **Results**
Advanced benchmark harness

- User
- Datasets
  - Dataset Generator
- Configuration
- Benchmark Core
  - Platform-specific algorithm implementation
- Report Generator
- Output Validator
- System Monitor
- Results
- Graph processing platform
Advanced benchmark harness

User

Datasets

Dataset Generator

Benchmark Core

Platform-specific algorithm implementation

Configuration

Report Generator

Output Validator

Graph processing platform

System Monitor

Results
Advanced benchmark harness

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

System Monitor

Results

Graph processing platform
Advanced benchmark harness

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

Results

System Monitor

Graph processing platform

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

Results

System Monitor

Graph processing platform
Advanced benchmark harness

- Configuration
- Benchmark Core
- Platform-specific algorithm implementation
- Output Validator
- Report Generator
- Results
- System Monitor
- Graph processing platform
- Dataset Generator
- Datasets
- User
Advanced benchmark harness
Advanced benchmark harness

User

Datasets

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

System Monitor

Results

Graph processing platform

Dataset Generator

Configuration

Benchmark Core

Platform-specific algorithm implementation

Report Generator

Output Validator

System Monitor

Results
## Supported algorithms

<table>
<thead>
<tr>
<th>ID</th>
<th>Algorithm</th>
<th>Class</th>
<th>Use (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFS</td>
<td>Breadth-first search</td>
<td>Traversal</td>
<td>46</td>
</tr>
<tr>
<td>STATS</td>
<td>Local clustering coefficient</td>
<td>Statistics</td>
<td>16</td>
</tr>
<tr>
<td>CONN</td>
<td>Weakly connected components</td>
<td>Connected components</td>
<td>13</td>
</tr>
<tr>
<td>CD</td>
<td>Label propagation</td>
<td>Community detection</td>
<td>5</td>
</tr>
<tr>
<td>EVO</td>
<td>Forest fire evolution</td>
<td>Evolution</td>
<td>4</td>
</tr>
</tbody>
</table>

Guo et al., How Well Do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis, IPDPS 2014
Experimental setup

- **DAS-4 cluster**
  - Typical big data setup
  - 11 nodes, 24 GiB RAM, 2 x 8-core Xeon E5620
  - 1 Gbit/s Ethernet

- **Single machine**
  - HPC-like setup
  - 192 GiB RAM, 2 x 8-core Xeon E5-2450 v2
Runtime

Graph500 23

Runtime (s)

10^2
10^2.5
10^3
10^3.5
10^4
51 120 8243 4647
Giraph GraphX MapReduce Neo4j

Patents

Runtime (s)

10^2
10^2.5
10^3
10^3.5
10^4
45 154 562 1228
Giraph GraphX MapReduce Neo4j

SNB 1000

Runtime (s)

10^2
10^2.5
10^3
10^3.5
10^4
71 178 2678
Giraph GraphX MapReduce Neo4j

Platform

TU Delft
Runtime

Graph500 23

8243
4647

Giraph
GraphX
MapReduce
Neo4j

Patents

1228
562

Platform

Giraph
GraphX
MapReduce
Neo4j

SNB 1000

2678

Giraph
GraphX
MapReduce
Neo4j

CONN

51
120

51
45
71

178

154

1228

562
Runtime

Graph500 23

Runtime (s)

Patents

Runtime (s)

SNB 1000

Runtime (s)

Platform

Giraph

GraphX

MapReduce

Neo4j

51

8243

120

4647

154

71

45

1228

562

178

2678

TUDelft
Runtime

2 orders of magnitude difference
Neo4j 2x faster than MR
Runtime

MR 2x faster than Neo4j
Runtime

Graph500 23

Runtime (s)

GraphX
MapReduce
Neo4j

51
120
8243
4647

Giraph
MapReduce
Neo4j

154
1228
562

Patents

SNB 1000

Runtime (s)

CONN

Giraph
MapReduce
Neo4j

71
178
2678

Giraph
GraphX
MapReduce
Neo4j

45
45
Edge-normalized performance

Throughput (kEPS)

Graph500 23

Patents

SNB 1000

Platform

Giraph
GraphX
MapReduce
Neo4j

Giraph
GraphX
MapReduce
Neo4j

Giraph
GraphX
MapReduce
Neo4j

CONN

TU Delft
Edge-normalized performance

Throughput (kEPS)

Graph500 23

Patents

SNB 1000

Number of edges in graph / runtime

Platform

Giraph
GraphX
MapReduce
Neo4j
Edge-normalized performance

20 x difference because of dataset structure
Edge-normalized performance

Also for other platforms
Edge-normalized performance

Graph500 23

Patents

SNB 1000

Throughput (kEPS)

Platform

Giraph
GraphX
MapReduce
Neo4j

Giraph
GraphX
MapReduce
Neo4j

Giraph
GraphX
MapReduce
Neo4j

CONN

2551 1081

364 108

6272 2508

16 28

13

167
See paper for more results
Conclusion
Conclusion

- Use Graphalytics to:  
  - Compare  
  - Tune  
  - Re-design
Conclusion

- Use Graphalytics to:
  - Compare
  - Tune
  - Re-design

- Open source (Apache License 2.0)
  - Contribute an implementation for your platform
Conclusion

• Use Graphalytics to:
  – Compare
  – Tune
  – Re-design

• Open source (Apache License 2.0)
  – Contribute an implementation for your platform

• SPEC RG standardization coming
Performance evaluation – Granular

BSP Programming Model
Deploying Granular

Performance Model

Patch

Analyzer

BDP System

Granular Archiver

Granular Archive

Performance Modeling

Performance Archiving

TU Delft
Granular results

Network Traffic

Execution Time (s)

Message Volume (M)

TS-SentMsgVolume

TS-LocalMsgVolume

TS-RemoteMsgVolume
Thank you!

mihai@mihaiic.ro

graphalytics.ewi.tudelft.nl

GitHub

github.com/tudelft-atlarge/graphalytics