

# ***Social Network Benchmark: Business Intelligence workload***

Gábor Szárnyas

[szarnyas@db.bme.hu](mailto:szarnyas@db.bme.hu)

10th LDBC TUC meeting

1 September, Munich



# *The SNB task force*



Arnau Prat  
Sparsity / DAMA-UPC  
(Task Force Leader)



Marcus Paradies  
SAP



Moritz Kaufmann  
TUM / Tableau



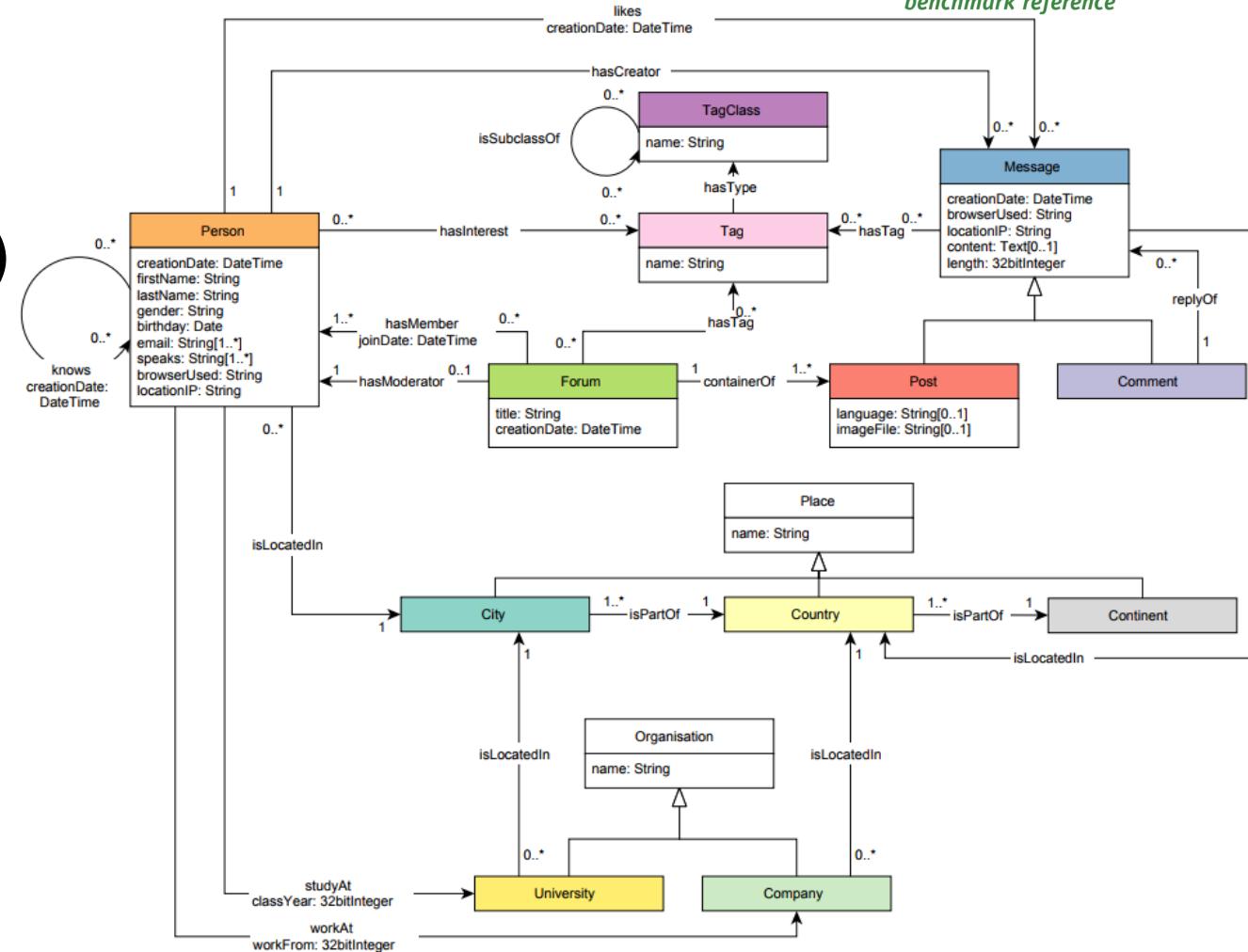
Alex Averbuch  
Neo4j



Gábor Szárnyas  
BME / McGill

# Social Network Benchmark

- Social network graph
- Scalable generator (DATAGEN)
- Analytical workload
  - Graphalytics (VLDB 2016)
- Query evaluation workloads
  - Interactive (SIGMOD 2015)
  - Business Intelligence
- Choke-point driven design



# Challenges



# Queries difficult to comprehend

## Complex Reads Query Descriptions

Notes:

- Some queries require returning the content of a post. As stated in the schema, posts have content or an imageFile, but not both. An empty string in content represents the post not having content, therefore, it must have a non empty string in imageFile and the other way around.

### 1. Friends with certain name

- **Description:** Given a start Person, find Persons with a given first name that the start Person is connected to (excluding start Person) by at most 3 steps via Knows relationships. Return Persons, including summaries of the Persons workplaces and places of study.

- **Parameters:**

Person.id      ID

Person.firstName      String

- **Results:**

Person.id      ID

Person.lastName      String

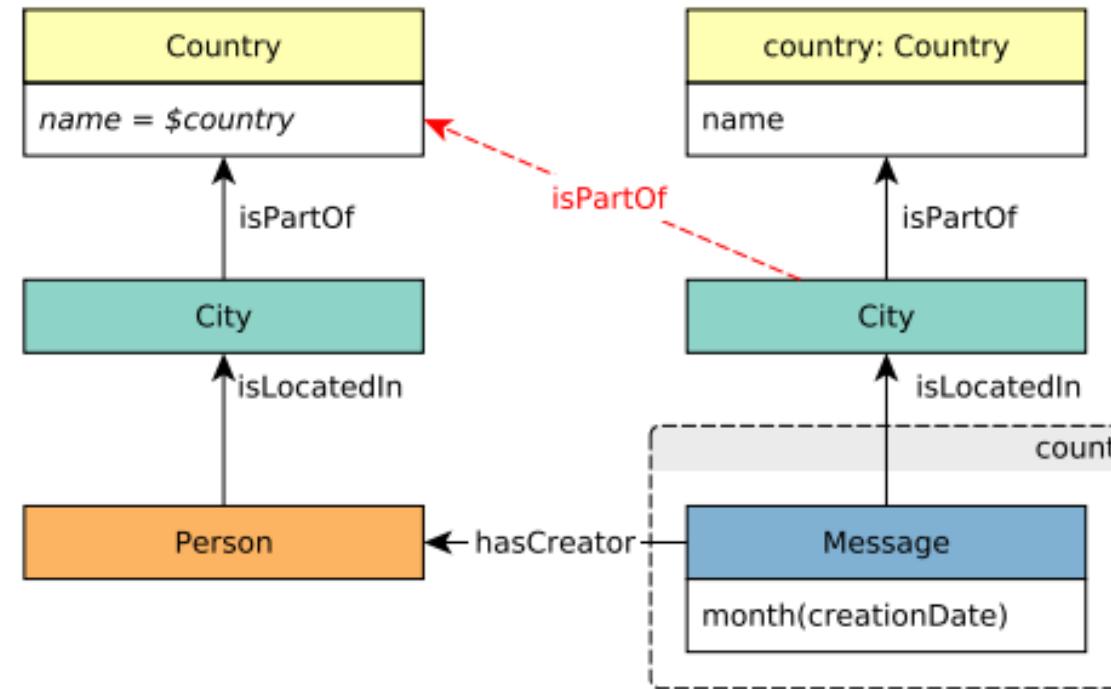


# *Outdated docs / inconsistencies*

APP  
post

3 Post.id	ID
4 Post.imageFile	String
5 Post.creationDate	DateTime
6 Post.locationIp	String
7 Post.browserUsed	String
8 Post.language	String
9 Post.content	Text
10 Post.length	32-bit Integer
11 Post-hasCreator->Person.id	ID
12 Forum-containerOf->Post	ID
13 Post-isLocatedIn->Country.id	ID
14 {Post-hasTag->Tag.id}	{ID}

# Graphical notation

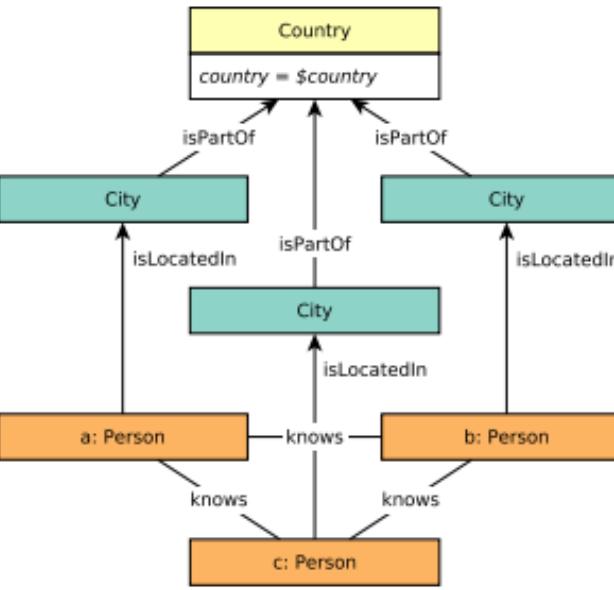


# ***Query specification***

- YAML files as a single source of truth
  - Generating query cards (TeX)
  - Generating wiki entries (Markdown) is also possible
- Python/Jinja2 for defining templates

```
1 workload: BI
2 operation: read
3 number: 17
4 title: Friend triangles
5 description: |
6   For a given country, count all the distinct triples of persons such that `a` is friend of `b`, `b` is friend of `c`,
7   and `c` is friend of `a`.
8
9 Distinct means that given a triple $t_1$ in the result set $R$ of all qualified triples,
10 there is not a triple $t_2$ in $R$ such that $t_1$ and $t_2$ have the same set of elements.
11 parameters:
12   - name: country
13     type: String
14 result:
15   - name: count
16     type: 32-bit Integer
17     category: aggregated
18 choke_points: [1.1, 2.3]
```

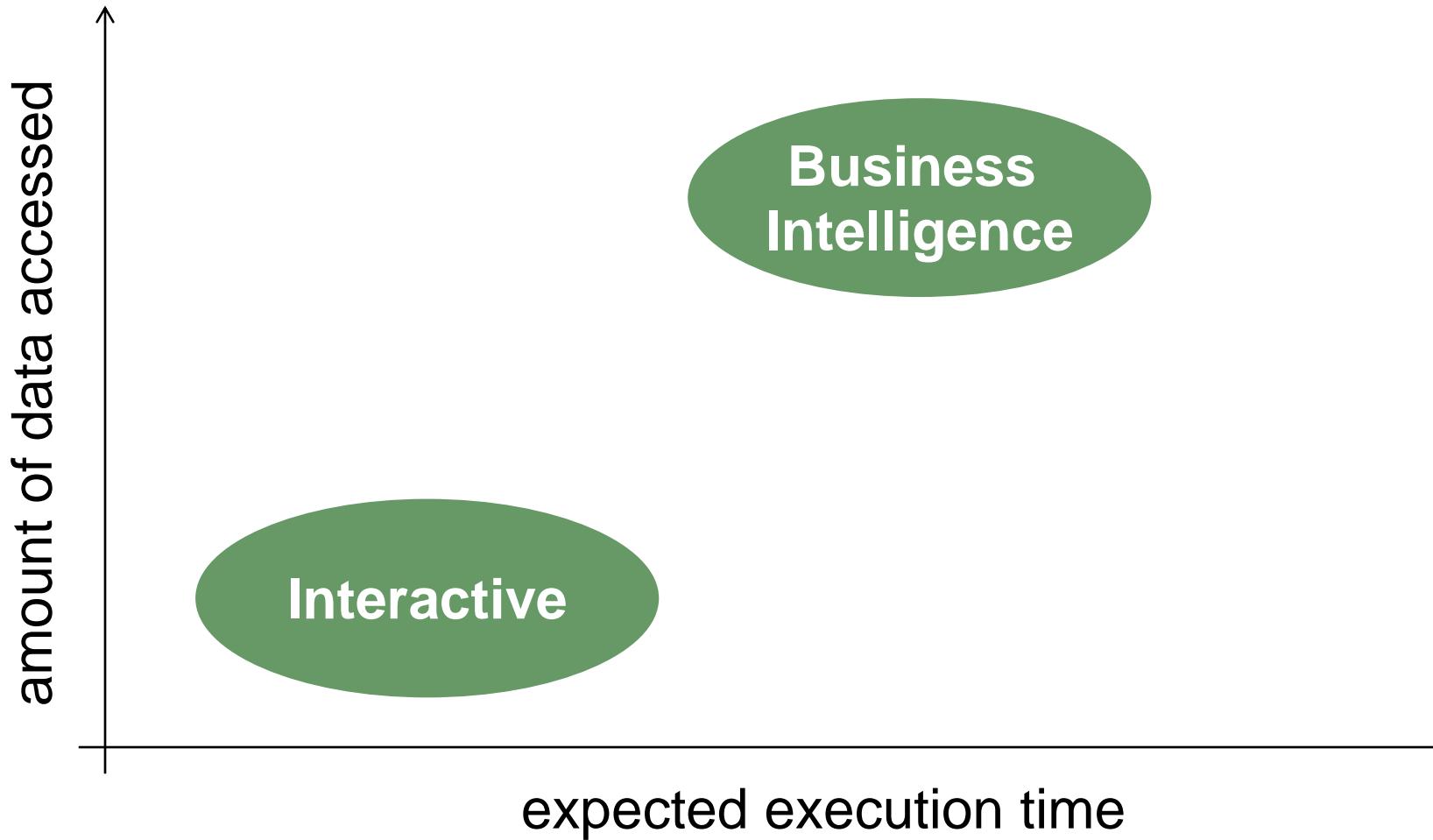


query	BI / read / 17					
title	Friend triangles					
pattern	 <pre>     graph TD       Country[Country&lt;br/&gt;country = \$country] --&gt; City1[City]       Country --&gt; City2[City]       Country --&gt; City3[City]              City1 -- isPartOf --&gt; Country       City2 -- isPartOf --&gt; Country       City3 -- isPartOf --&gt; Country              City1 -- isLocatedIn --&gt; a[a: Person]       City1 -- isLocatedIn --&gt; b[b: Person]              City2 -- isLocatedIn --&gt; a       City2 -- isLocatedIn --&gt; c[c: Person]              City3 -- isLocatedIn --&gt; b       City3 -- isLocatedIn --&gt; c              a -- knows --&gt; b       b -- knows --&gt; c       b -- knows --&gt; a   </pre>					
desc.	<p>For a given country, count all the distinct triples of persons such that a is friend of b, b is friend of c, and c is friend of a.</p> <p>Distinct means that given a triple <math>t_1</math> in the result set <math>R</math> of all qualified triples, there is not a triple <math>t_2</math> in <math>R</math> such that <math>t_1</math> and <math>t_2</math> have the same set of elements.</p>					
params	<table border="1"> <tr> <td>1</td> <td>country</td> <td>String</td> <td></td> </tr> </table>	1	country	String		
1	country	String				
result	<table border="1"> <tr> <td>1</td> <td>count</td> <td>32-bit Integer</td> <td>R</td> <td></td> </tr> </table>	1	count	32-bit Integer	R	
1	count	32-bit Integer	R			
CPs	1.1, 2.3					

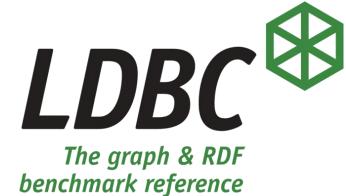
# Query evaluation workloads



# *Query evaluation workloads*



# Interactive workload



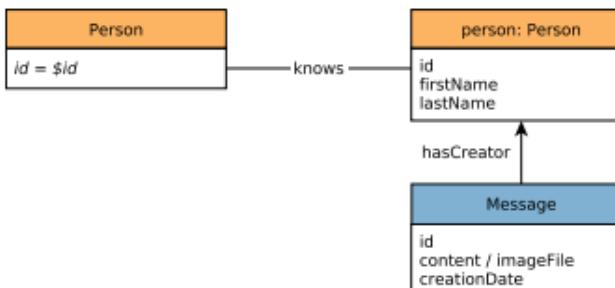
- 14 complex read queries
- 8 short read queries
- 7 update queries
- **Queries explore the graph around a given node**

O. Erling (Openlink), A. Averbuch (Neo), J.L. LarribaPey (UPC), Hassan Chafi (Oracle Labs), Andrey Gubichev (TU Munich), Arnau Prat (DAMA-UPC), Minh-Duc Pham (VU Amsterdam), Peter Boncz (CWI).

**The LDBC Social Network Benchmark: Interactive Workload.**

Proceedings of SIGMOD 2015, Melbourne



query	Interactive / complex / 2						
title	Recent posts and comments by your friends						
pattern							
desc.	Given a start Person, find (most recent) Messages from all of that Person's friends, that were created before (and including) a given date.						
params	1	Person.id	ID				
	2	date	DateTime				
result	1	Message-hasCreator->Person.id	ID	R			
	2	Message-hasCreator->Person.firstName	String	R			
	3	Message-hasCreator->Person.lastName	String	R			
	4	Message.id	ID	R			
	5	Message.content or Post.imageFile	String	R			
	6	Message.creationDate	DateTime	R			
sort	1	Message.creationDate	↓				
	2	Message.id	↑				
limit	20						
CPs	1.1, 2.2, 2.3, 3.2						

# Interactive workload

1

LDBC Social Network Benchmark (SNB) - 0.3.0

2

Page 28 of 86 LDBC Social Network Benchmark (SNB)

3

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

4

Page 28 of 86 LDBC Social Network Benchmark (SNB)

5

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

6

Page 30 of 86 LDBC Social Network Benchmark (SNB)

7

LDBC Social Network Benchmark (SNB) - 0.3.0 Page 31 of 86

8

LDBC Social Network Benchmark (SNB) - 0.3.0 Page 32 of 86

9

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

10

LDBC Social Network Benchmark (SNB) - 0.3.0 Page 34 of 86

11

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

12

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

13

LDBC Social Network Benchmark (SNB) - 0.3.0 Page 38 of 86

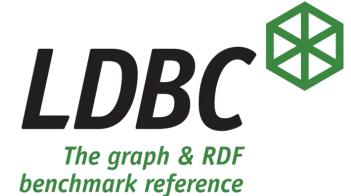
14

LDBC Social Network Benchmark (SNB) - 0.3.0 Page

15

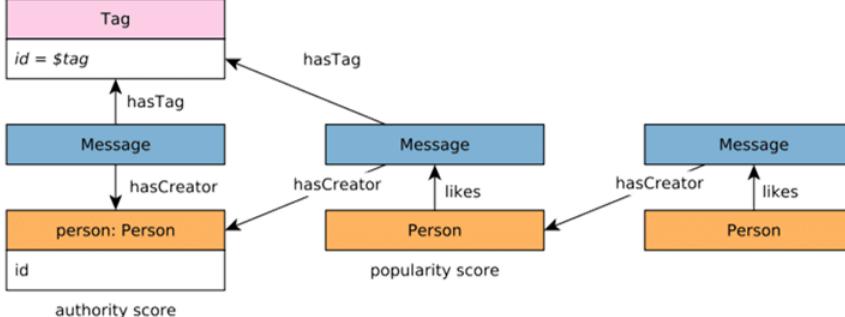


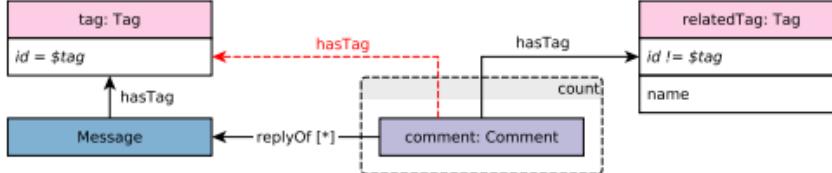
# *Business Intelligence workload*



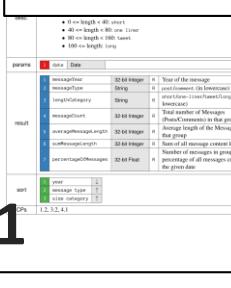
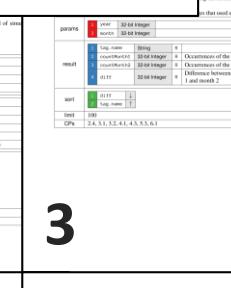
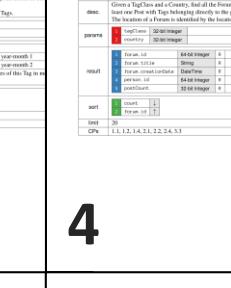
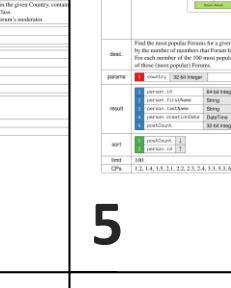
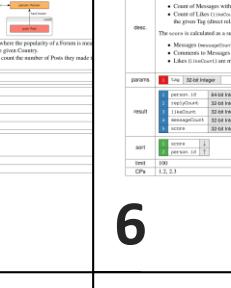
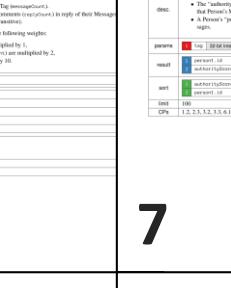
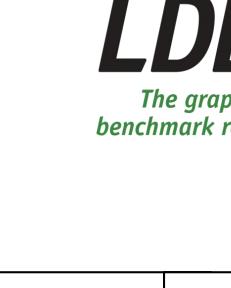
- 25 read queries
- Batch updates (not yet defined)
- **Queries explore vast portions of the graph**



query	BI / read / 7										
title	Most authoritative users on a given topic										
pattern	 <p style="text-align: center;">authority score</p>										
desc.	<p>Given a Tag, find all Persons that ever created a Message with the given Tag. For each of these Persons compute their “authority score” as follows:</p> <ul style="list-style-type: none"> <li>• The “authority score” is the sum of “popularity scores” of the Persons that liked any of that Person’s Messages with the given Tag.</li> <li>• A Person’s “popularity score” is defined as the total number of likes on all of their Messages.</li> </ul>										
params	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="background-color: red; color: white; padding: 2px 5px;">1</td> <td>tag</td> <td>32-bit Integer</td> <td colspan="2"></td> </tr> </table>	1	tag	32-bit Integer							
1	tag	32-bit Integer									
result	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="background-color: blue; color: white; padding: 2px 5px;">1</td> <td>person1.id</td> <td>64-bit Integer</td> <td>R</td> <td></td> </tr> <tr> <td style="background-color: blue; color: white; padding: 2px 5px;">2</td> <td>authorityScore</td> <td>32-bit Integer</td> <td>R</td> <td></td> </tr> </table>	1	person1.id	64-bit Integer	R		2	authorityScore	32-bit Integer	R	
1	person1.id	64-bit Integer	R								
2	authorityScore	32-bit Integer	R								
sort	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="background-color: green; color: white; padding: 2px 5px;">1</td> <td>authorityScore</td> <td style="text-align: center;"><math>\downarrow</math></td> </tr> <tr> <td style="background-color: green; color: white; padding: 2px 5px;">2</td> <td>person1.id</td> <td style="text-align: center;"><math>\uparrow</math></td> </tr> </table>	1	authorityScore	$\downarrow$	2	person1.id	$\uparrow$				
1	authorityScore	$\downarrow$									
2	person1.id	$\uparrow$									
limit	100										
CPs	1.2, 2.3, 3.2, 3.3, 6.1										

query	BI / read / 8										
title	Related topics										
pattern											
desc.	Find all Messages that have a given Tag. Find the related Tags attached to replies of these Messages (direct relation not transitive). but only of those replies that do not have the given Tag. Group the Tags by name, and get the count of replies in each group.										
params	<table border="1"> <tr> <td>1</td> <td>tag</td> <td>32-bit Integer</td> <td></td> </tr> </table>	1	tag	32-bit Integer							
1	tag	32-bit Integer									
result	<table border="1"> <tr> <td>1</td> <td>relatedTag.name</td> <td>String</td> <td>R</td> <td></td> </tr> <tr> <td>2</td> <td>count</td> <td>32-bit Integer</td> <td>R</td> <td></td> </tr> </table>	1	relatedTag.name	String	R		2	count	32-bit Integer	R	
1	relatedTag.name	String	R								
2	count	32-bit Integer	R								
sort	<table border="1"> <tr> <td>1</td> <td>count</td> <td>↓</td> </tr> <tr> <td>2</td> <td>relatedTag.name</td> <td>↑</td> </tr> </table>	1	count	↓	2	relatedTag.name	↑				
1	count	↓									
2	relatedTag.name	↑									
limit	100										
CPs	1.6, 3.3, 5.2										

# BI workload

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25				
																												
<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>	<b>Chapter 5: Business Intelligence Workload</b>				
<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>	<b>5.1. Read Query Descriptions</b>				
<b>BI / read / 1</b> query: BI / read / 1 title: Friend triangles	<b>BI / read / 2</b> query: BI / read / 2 title: How many persons have a given number of posts	<b>BI / read / 3</b> query: BI / read / 3 title: Stranger's interaction	<b>BI / read / 4</b> query: BI / read / 4 title: Popular topics in a country	<b>BI / read / 5</b> query: BI / read / 5 title: Top posters in a country	<b>BI / read / 6</b> query: BI / read / 6 title: Most active posters of a given Topic	<b>BI / read / 7</b> query: BI / read / 7 title: Most active posters on a given topic	<b>BI / read / 8</b> query: BI / read / 8 title: Related topics	<b>BI / read / 9</b> query: BI / read / 9 title: Forum with related Tags	<b>BI / read / 10</b> query: BI / read / 10 title: Create Person for a Tag	<b>BI / read / 11</b> query: BI / read / 11 title: Unrelated replies	<b>BI / read / 12</b> query: BI / read / 12 title: Popular Posts in a month	<b>BI / read / 13</b> query: BI / read / 13 title: Popular Tags in a month	<b>BI / read / 14</b> query: BI / read / 14 title: Social normals	<b>BI / read / 15</b> query: BI / read / 15 title: Social normals	<b>BI / read / 16</b> query: BI / read / 16 title: Experts in social circle	<b>BI / read / 17</b> query: BI / read / 17 title: Friend triangles	<b>BI / read / 18</b> query: BI / read / 18 title: Stranger's interaction	<b>BI / read / 19</b> query: BI / read / 19 title: High-level topics	<b>BI / read / 20</b> query: BI / read / 20 title: Zombies in a country	<b>BI / read / 21</b> query: BI / read / 21 title: International dialog	<b>BI / read / 22</b> query: BI / read / 22 title: Holiday destinations	<b>BI / read / 23</b> query: BI / read / 23 title: Messages by Topic and Content	<b>BI / read / 24</b> query: BI / read / 24 title: Weighted paths	<b>BI / read / 25</b> query: BI / read / 25 title: Weighted paths				
<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>	<b>params:</b> <code>pattern</code>					
<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>	<b>result:</b> <code>result</code>				
<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>	<b>1CPs:</b> <code>1CPs</code>				

# Choke points



## A.2 Join Performance

### CP-2.1: [QOPT] Rich join order optimization

TPC-H 2.3

This choke-point tests the ability of the query optimizer to find optimal join orders. A graph can be traversed in different ways. In the relational model, this is equivalent as different join orders. The execution time of these orders may differ by orders of magnitude. Therefore, finding an efficient join (traversal) order is important, which in general, requires enumeration of all the possibilities. The enumeration is complicated by operators that are not freely re-orderable like semi-, anti-, and outer-joins. Because of this difficulty most join enumeration algorithms do not enumerate all possible plans, and therefore can miss the optimal join order. Therefore, these chokepoint tests the ability of the query optimizer to find optimal join (traversal) orders.

**Queries.** BI 2 BI 4 BI 5 BI 9 BI 10 BI 11 BI 19 BI 20 BI 21 BI 22 BI 24 BI 25  
Interactive 1 Interactive 3

TPC-H Analyzed: Hidden Messages and Lessons Learned from an Influential Benchmark

Peter Boncz, Thomas Neumann, Orri Erling.

TPC-H Analyzed: Hidden Messages and Lessons Learned from an Influential Benchmark.

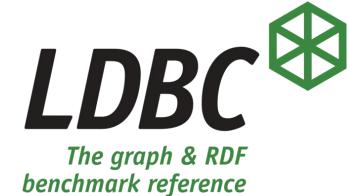
TPCTC 2013



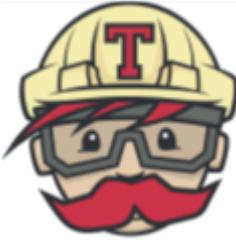
	1.1	1.2	1.3	1.4	1.5	1.6	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	6.1	7.1	7.2	7.3	7.4	
BI 1			●								●			●											
BI 2	●	●		●			●		●		●	●													
BI 3									●		●	●				●				●		●			
BI 4	●	●		●			●	●		●				●											
BI 5		●		●	●		●	●	●		●									●		●			
BI 6			●						●																
BI 7		●							●			●	●									●			
BI 8						●							●								●				
BI 9		●		●			●		●		●														
BI 10		●					●		●				●												
BI 11	●						●	●	●		●	●										●			
BI 12		●						●				●									●				
BI 13		●						●	●				●								●				
BI 14		●						●	●				●								●	●	●	●	
BI 15		●							●				●	●						●	●				
BI 16		●			●				●			●									●	●	●	●	
BI 17	●								●																
BI 18		●	●			●							●			●	●	●							
BI 19		●		●			●		●	●				●				●				●	●	●	
BI 20						●		●		●										●					
BI 21		●					●		●	●			●	●				●		●					
BI 22			●		●		●		●			●	●					●	●	●					
BI 23				●				●	●			●					●								
BI 24				●		●		●	●			●	●				●				●				
BI 25		●					●	●	●			●						●			●		●	●	



# *Continuous integration*

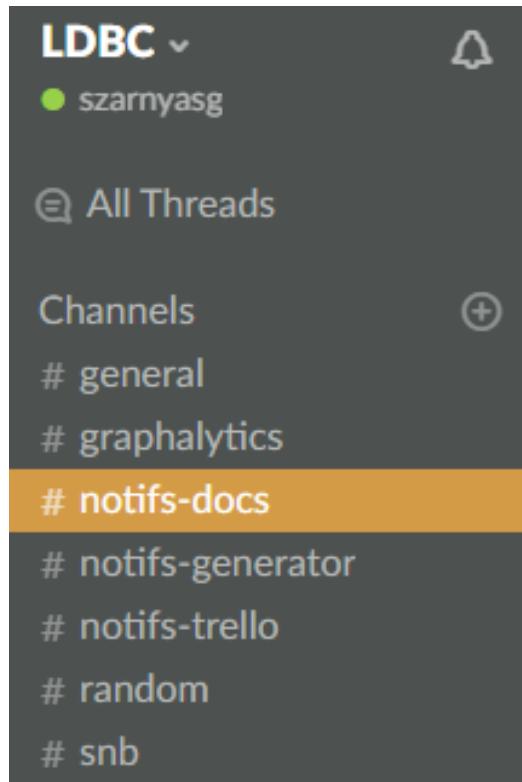


- Use Travis CI to
  - build DATAGEN
  - generate the technical report
- [https://github.com/LDBC/LDBC\\_snb\\_docs](https://github.com/LDBC/LDBC_snb_docs)
- [https://github.com/LDBC/LDBC\\_snb\\_datagen](https://github.com/LDBC/LDBC_snb_datagen)



# LDBC Slack team

<https://ldbcouncil.slack.com/>



#notifs-docs

☆ | 8 5 | 0 | [http://ldbc.github.io/ldbc\\_snb\\_docs/](http://ldbc.github.io/ldbc_snb_docs/)...

Wednesday, August 16th

github APP 3:05 AM  
[ldbc\_snb\_docs:dev] 1 new commit by Gabor Szarnyas:  
5d94533 Cleanup in docs and for CPs in particular - Gabor Szarnyas

github APP 3:11 AM  
[ldbc\_snb\_docs:dev] 1 new commit by Gabor Szarnyas:  
dfb73a6 Remove list of figures and list of tables, fix section names - Gabor Szarnyas

[ldbc\_snb\_docs:dev] 1 new commit by Gabor Szarnyas:  
2c9d71e Fix refs - Gabor Szarnyas



# Progress

- 54 Trello cards
- Documentation
  - 180+ commits
  - 14 issues
  - +12 LaTeX packages
- DATAGEN
  - 40+ commits
  - Talk by Arnau at 13:30

# Roadmap

- Implement & validate for Neo4j, PostgreSQL and Sparksee
- Publish a subset of the benchmark in a workshop
  - GraphQ @ EDBT (late Nov)
  - GRADES @ SIGMOD (late March)
- Gather feedback & refine
- Define update operations
- We are recruiting!

