LDBC

The graph & RDF benchmark reference

LDBC SNB Business Intelligence Workload: Chokepoint Analysis

Arnau Prat

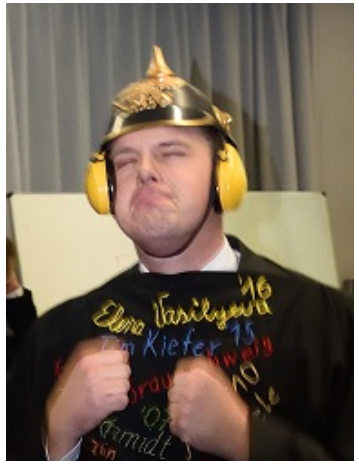
9th LBDC TUC meeting

9/10 February Walldorf (Germany)

The Team

LDBC
The graph & RDF benchmark reference

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



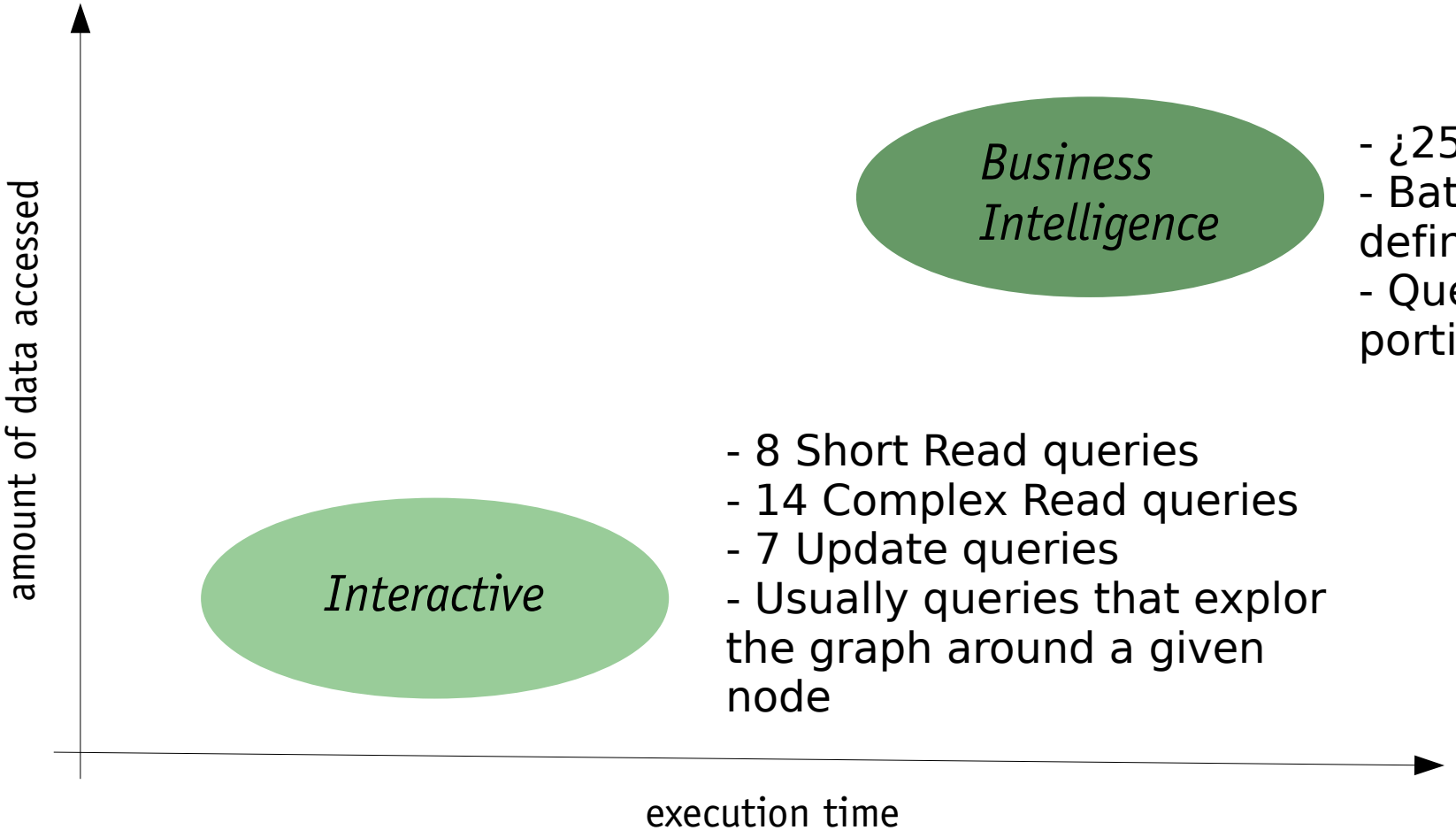
Marcus Paradies
SAP
(Minion)

Moritz Kaufmann
TUM
(Minion)

Alex Averbuch
Neo
(Minion)

A bit of recap

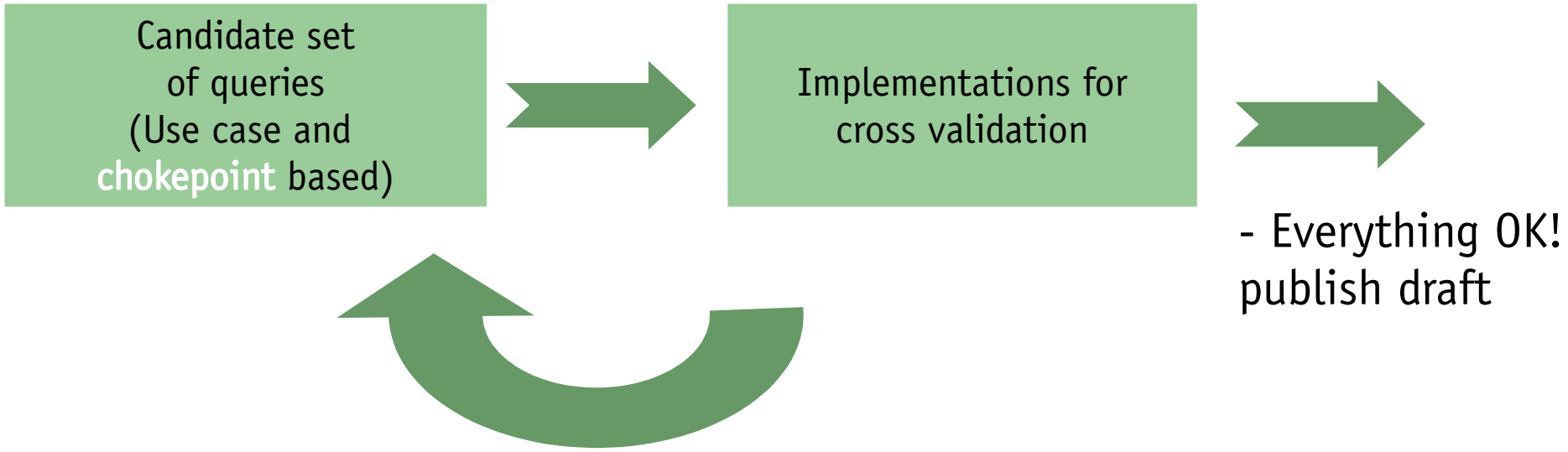
LDBC
The graph & RDF benchmark reference



- ¿25? BI queries
- Batch updates (not yet defined)
- Queries explore vast portions of the graph

- 8 Short Read queries
- 14 Complex Read queries
- 7 Update queries
- Usually queries that explor the graph around a given node

The LDBC SNB Algorithm



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
CP1.1		*							*					
CP1.2									*					
CP1.3	*													
CP1.4								*			*			
CP1.5												*		
CP2.1	*		*											
CP2.2		*					*		*					
CP2.3		*		*	*		*		*	*	*			
CP2.4								*			*			
CP3.1			*											
CP3.2		*						*	*					
CP3.3					*		*	*	*	*	*	*	*	*
CP4.1										*				
CP4.2										*				
CP5.1			*			*	*			*				
CP5.2										*				
CP5.3	*													
CP6.1										*				
CP7.1										*				
CP7.2												*	*	*
CP7.3	*											*	*	*

Chokepoint coverage of Interactive Workload Complex Reads

The LDBC SNB Algorithm

LDBC
The graph & RDF benchmark reference

The LDBC SNB Algorithm.patch

LDBC
The graph & RDF benchmark reference

```
for ( Query q : biQueries ) {
    improveQueryFormulation(q);
    findChokepoints(q);
}

addRemoveNewQueriesIfNecessary();
publishDraft();
```

Chokepoint Analysis - BI Workload

LDBC
The graph & RDF benchmark reference

24 BI queries:

1 Posting summary
2 Top tags for country, age, gender, time
3 Tag evolution
4 Popular topics in a country
5 Top posters in a country
6 Most active posters of a given topic
7 Most authoritative users on a given topic
8 Related topics
9 Forums with related tags
10 Central Person for a tag
11 Unrelated replies
12 Trending posts
13 Popular tags per month in a country
14 Top thread initiators
15 Social normals
16 Experts in social circle
17 Friend triangles
18 Persons with a given number of posts
19 Strangers interaction
20 High level topics
21 Zombies in a country
22 International dialog
23 Holiday destinations
24 Messages by topic and continent

Chokepoint Analysis - Query 14 Top Thread Initiators

LDBC
The graph & RDF benchmark reference

Description
For each person, count the number Posts they created in the time interval (begin,end), and the number of messages in each of their (transitive) reply trees.
When calculating message counts only consider messages created within the given time interval.
Return each person, number of Posts they created, and the count of all messages that appeared in the reply trees (including Post at tree root) they created.

Parameters:
begin - Date
end - Date

Result:
Person.id - 64-bit Integer
Person.first_name - String
Person.last_name - String
thread_count - 32-bit Integer // The number of threads initiated by that Person
message_count - 32-bit Integer // The number of messages created in all the threads this Person initiated

Sort:
1st message_count (descending)
2nd Person.id (ascending)

Limit:
100

Chokepoint Analysis

LDBC
The graph & RDF benchmark reference

- TPC-H Analyzed: Hidden Messages and Lessons Learned from an Influential Benchmark - Peter Boncz, Thomas Neumann and Orri Erling - TPCTC 2013

- Chokepoints from LDBC Interactive Workload Deliverable D3.3.34 (http://ldbcouncil.org/sites/default/files/LDBC_D3.3.34.pdf)

Chokepoint Analysis

LDBC
The graph & RDF benchmark reference

- 1 Aggregation performance
  - 1.1 Interesting orders
  - 1.2 High Cardinality group-by performance
  - 1.3 Complex aggregate performance
  - 1.4 Top-k push-down
  - 1.5 Dependant group-by keys
  - 1.6 Low Cardinality group-by performance
- 2 Join performance
  - 2.1 Rich join order optimization
  - 2.2 Late projection
  - 2.3 Join type selection
  - 2.4 Sparse foreign keys
- 3 Data access locality
  - 3.1 Detecting correlation
  - 3.2 Dimensional clustering
  - 3.3 Scattered Index access patterns
- 4 Expression calculation
  - 4.1 Common subexpression elimination
  - 4.2 Complex boolean expressions
  - 4.3 Low overhead expressions interpretation
- 5 Sub-queries
  - 5.1 Flattening sub-queries
  - 5.2 Overlapp between outer and sub-query
  - 5.3 Intra-query result reuse
- 6 Parallelism and Concurrency
  - 6.1 Inter-query result reuse
- 7 RDF and Graph Specifics
  - 7.1 Translation of internal to external ids
  - 7.2 Cardinality estimation of transitive paths
  - 7.3 Execution of transitive step
  - 7.4 Efficient evaluation of termination criteria of transitive queries
  - 7-5 Path pattern reuse

Data Analytics 101

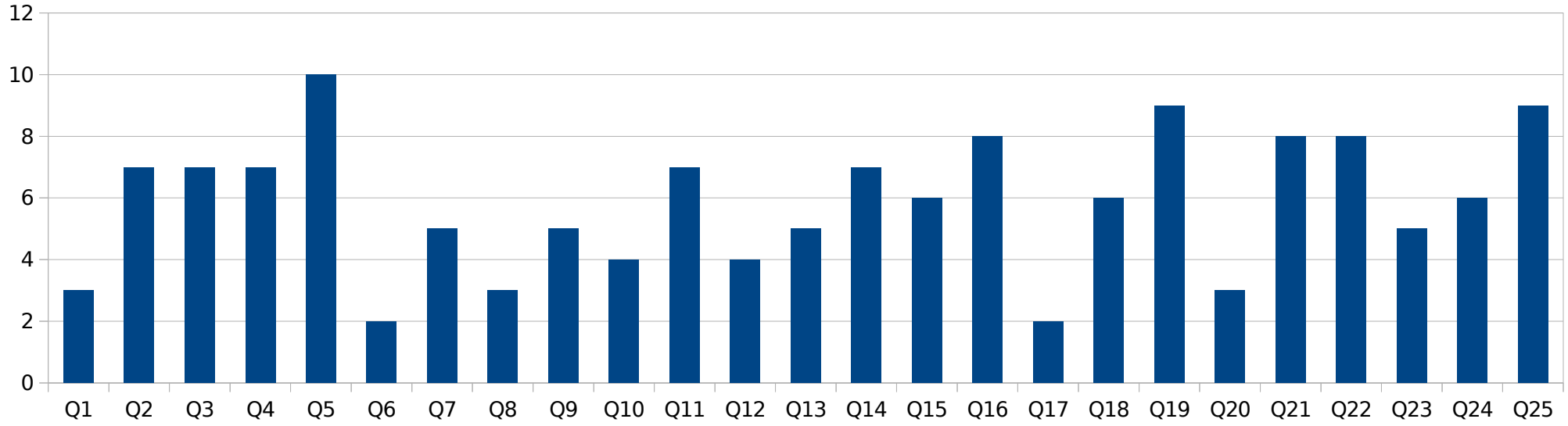
LDBC
The graph & RDF benchmark reference

- It took 35 1-hour meetings to do the analysis
- 146 Chokepoints in total
- This means:
  - 4.17 Chokepoints per hour
  - 0.7 queries per hour
- Added 2 new Chokepoints
- Changed the definition/formulation of 10 queries
- Added new query 25 Weighted Paths (an extension to interactive query 14)

Data Analytics 101

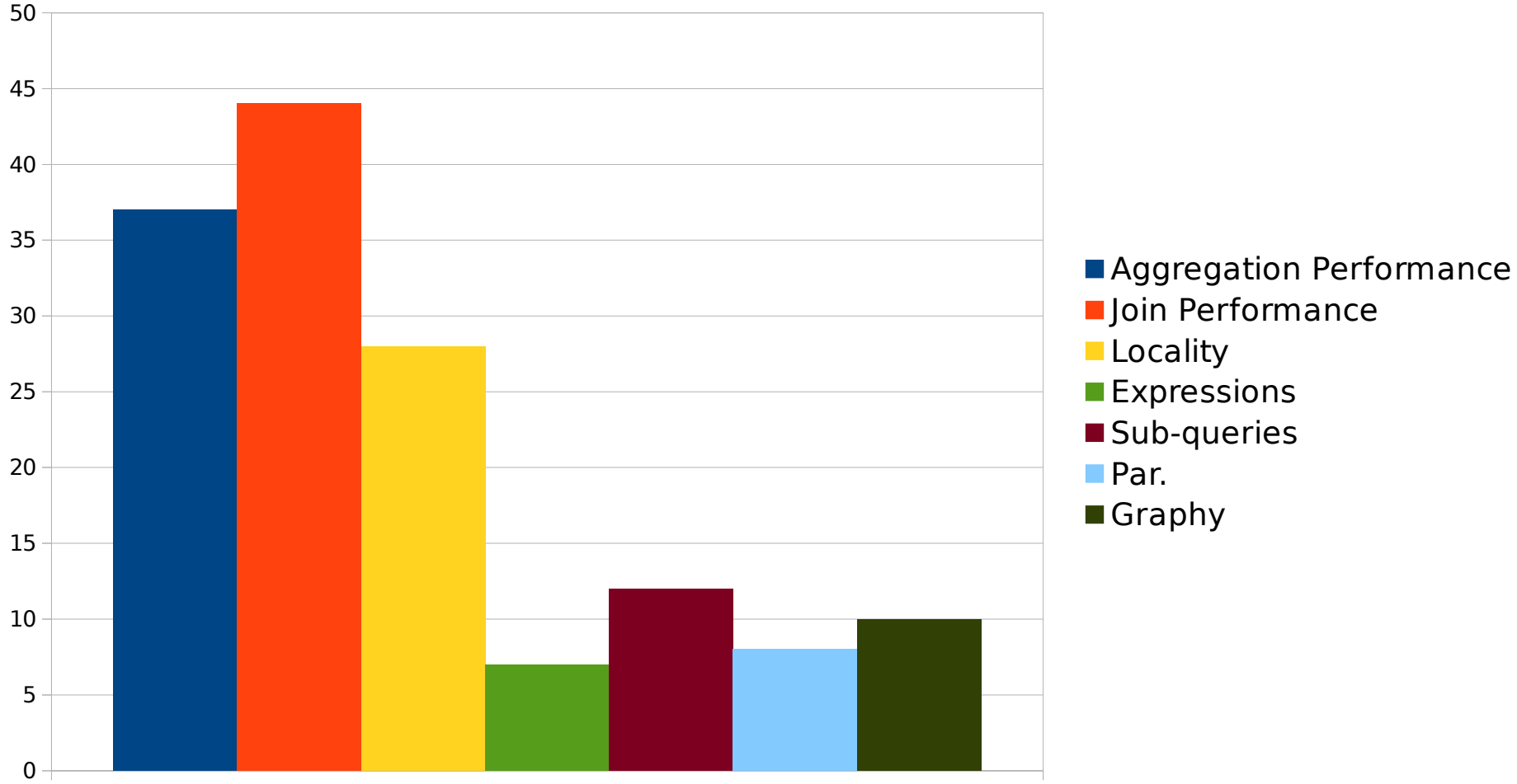
LDBC
The graph & RDF benchmark reference

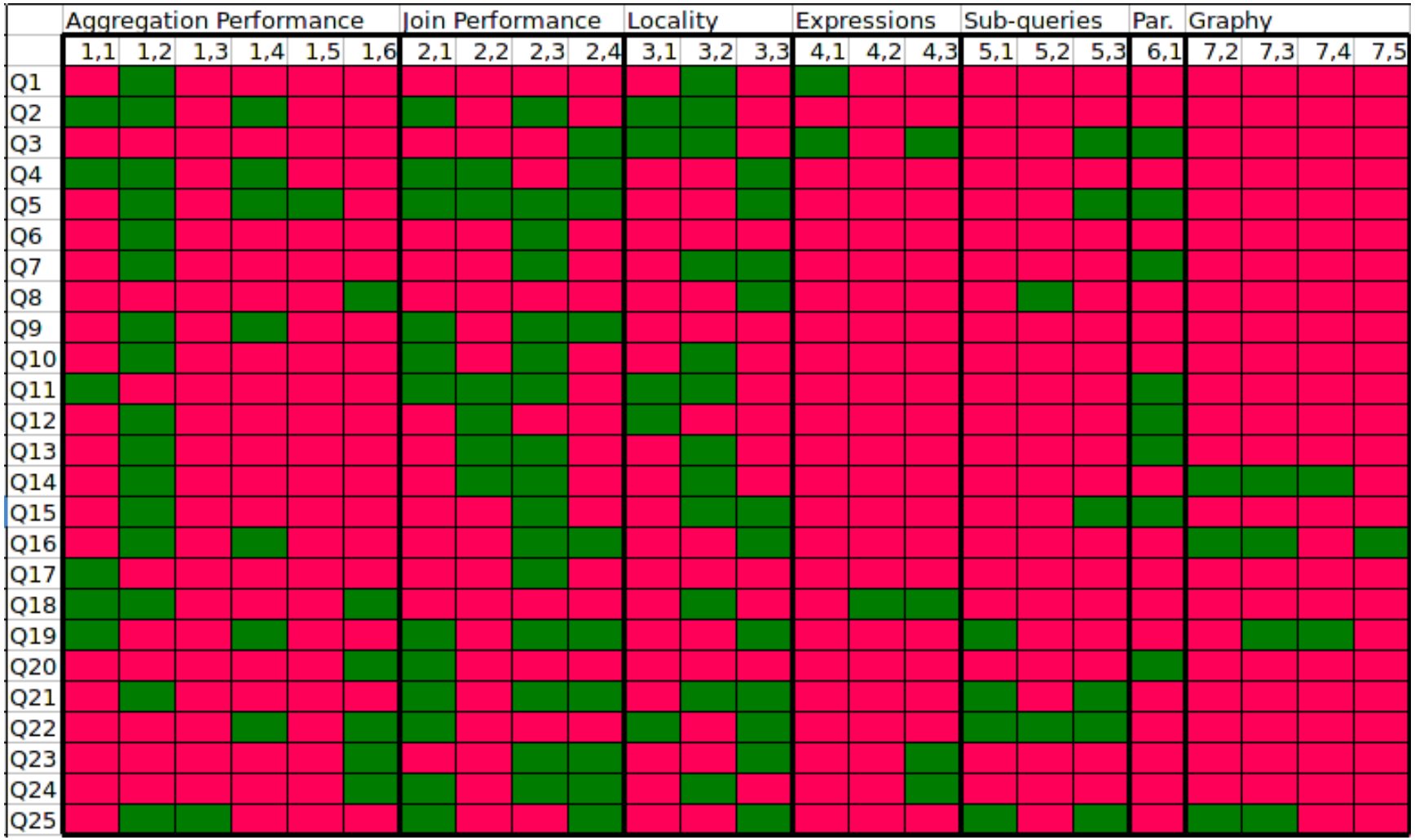
- Most fulfilled Chokepoints (16): 1.2-High Cardinality group by and 2.3-Join type selection
- Least fulfilled Chokepoints (1): 1.3-Complex aggregate performance, 1.5-Dependant group-by keys, 4.2-Complex boolean expressions and 7.5 - Path pattern reuse
- About 6 chokepoints per query in average. max is 10, min is 2.



Data Analytics 101

LDBC
The graph & RDF benchmark reference





LDBC
The graph & RDF benchmark reference

Data Analytics 101

LDBC
The graph & RDF benchmark reference

- We'd like to reduce the numer of queries to something around 20-22
- We've computed the “all-pairs Jaccard Coefficient” between the chokepoints of each query

Query	Query	Jaccard
7	15	0,83
7	13	0,67
2	10	0,57
13	15	0,57
23	24	0,57
2	11	0,56
15	21	0,56
4	5	0,55
21	25	0,55
2	9	0,5

Next steps

LDBC
The graph & RDF benchmark reference

- Take a decision on reducing the size of the workload
- Provide the validation datasets for Neo, PostgreSQL and Sparksee
- Introduce Batch Updates
- Define que query mix

LDBC
The graph & RDF benchmark reference

THANK YOU!

(and we are recruiting)