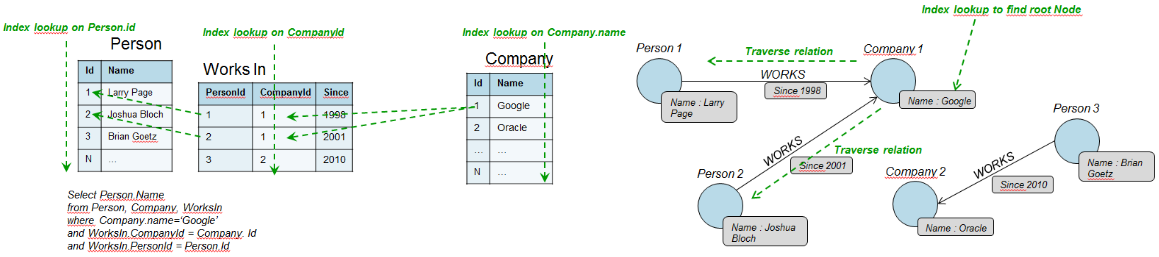
BDA+ Course, 2020

Graph vs Relational Databases



Person

Id	Name
1	Larry Page
2	Joshua Bloch
3	Brian Goetz
N	...

Works In

PersonId	CompanyId	Since
1	1	1998
2	1	2001
3	2	2010

Company

Id	Name
1	Google
2	Oracle
...	...
N	...

```
Select Person.Name
from Person, Company, WorksIn
where Company.name='Google'
and WorksIn.CompanyId = Company. Id
and WorksIn.PersonId = Person.Id
```

Context

In class we have seen, discussed and experimented with various types of databases to various lengths. One issue that needs to concern us is which database to use for which problem or in other words how do databases compare to each other. So for this small project you are asked to compare two (types of) databases, a graph database and a relational database respectively.

For the comparison to produce comparable and acceptable results you should make them over the same datasets and run them on the same hardware, having stopped or suspended all other programs that might be running at the same time. In your report you should provide information about the environment you used, e.g. machine configuration (CPU. Memory, etc.), Operating System, version of the corresponding software used and whatever else you find relevant.

Data

You are provided with aceess to online datasets. This years datasets that represent the Great Olympian Graph was chosen. The dataset is available at the class web site in four sheets. The Countries sheet including the country codes , the Cities sheet, including to the countries, the Sports Taxonomy sheet including the sports and disciplines and finally the Medallists sheet. One good tutorial to generate the neo4j graph model based on these four sheets is available in this.

You then need to extract the data from neo4j and import them to the relation database of your choice. You need to make sure that you load ***exactly*** the same data, otherwise your comparisons will not be valid. If for time or space issues you cannot load the full dataset please explain in your part which part you loaded and why.

Queries

You need to choose a set of at least 10 queries that you should express both in **SQL** and in **Cypher** and run those queries over the corresponding databases. You should choose queries that would show the advantages of both platforms. For example, these queries could include:

- Selection queries (e.g. get all unique athletes who have won a gold medal)
- Aggregation queries (e.g. find the country with the most gold medals)
- Path queries (e.g. find those athletes participating in Olympic Games with more than one country)

- Graph Metric calculation queries (e.g. find the most well connected athlete in the graph)
- Etc.

You are responsible for providing your own queries and explaining why you chose them. Note also that we expect to get the same number of results when we have two queries that are expected to answer the same question; if this is not the case you should recheck your queries or you should at least explain why this might be happening.

You can report your results in a form of table (like the one shown below) and/or the form of a graphic representation for a visual comparison of the results.

	Relational (time (ms) /number of results)	Graph (time (ms) /number of results)
Query 1	10 / 100	12 / 100
Query 2	123 / 1000	123.45 / 1000

Deliverables

You need to return a report describing your settings, the data you finally used for the experiments, the queries you chose and the results you have. You should also discuss in your report, besides the results you obtained, any kind of problems that you encountered during the process.

You should also return:

- A dump of the contents of your relational database as a zip file.
- The directory of the neo4j database as a zip file.