Query rewriting in the cloud

Location
Université Blaise Pascal – LIMOS, Clermont-Ferrand, France

Duration
5 or 6 months

Remuneration
Yes

Advisors
- Dr. Laurent D’Orazio, LIMOS, Université Blaise Pascal, Clermont-Ferrand, France (dorazio@isima.fr)
- Pr. Dominique Laurent, ETIS, Université de Cergy Pontoise, Cergy Pontoise, France (dominique.laurent@u-cergy.fr)
- Pr. Nicolas Spyratos, LRI, Université Paris Sud, Orsay, France (spyratos@lri.fr)

Context
In recent years, cloud computing has attracted increasing attention, especially under the leadership of suppliers such as Google, Microsoft, Amazon, Yahoo! or Facebook. Cloud computing aims at dealing with the growing needs in terms of computing and storage while being easily accessible. Particularly, clouds offer interesting perspectives for data analysis. Unfortunately, the performance of clouds are usually based on scalability relying on more computing nodes or more powerful ones, making the bill to pay higher for users and the use of resources far from optimal. Performance optimization in databases has been studied for decades of years. The use of methods such as indexing, materialized views or caches within the cloud would improve overall performance by optimizing resource utilization. In particular, materialized views and semantic caches can rewrite posed queries in order to reuse the results of previous ones. However, these solutions are based on languages (SQL, XPath, etc.) or well-known data models (relational, semi-structured, etc.). As a consequence, they are not suited for languages available in the cloud.

Objectives
The first objective of this work is to study in detail the evaluation of query analysis, including OLAP queries that use GROUP BY clauses and aggregate calculations by considering the different query languages data available in the cloud: Pig Latin, HiveQL, SCOPE or Jaql. Existing solutions must be extended to take into account the specificities these languages.
It will then be important to propose rewriting rules for analysis queries for them to respect the elasticity property of the cloud. In addition, it will be mandatory to consider specific operators that some languages include. The model will then be developed and validated using infrastructures such as the cluster of the Université Blaise Pascal, Amazon EC2 or Grid5000.

Plan
The different milestones are:
1. Study of the evaluation of analytical queries with the data models and the languages proposed in the cloud,
2. Proposal of a rewriting model for analytical queries to be applied to semantic caching and materialized views in the cloud relying on our previous work [5]
3. Validation of the model using private infrastructure (Amazon, etc.).

References