University of Zurich[UZH]

Department of Informatics

University of Zürich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zürich
Phone. +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch/dbtg

UZH, Dept. of Informatics, Binzmühlestr. 14, CH-8050 Zürich

Prof. Dr. Michael Böhlen
Professor
Phone +41 44 635 43 33
Fax +41 44 635 68 09
boehlen@ifi.uzh.ch

Zürich, March, 2017

BSc Thesis (18 KP)
Datenbanktechnologie

Topic: QR decomposition integration in MonetDB system

The demand to analyze data stored in DBMSs has increased significantly during the last few years. Since the analysis of scientific data is mostly based on statistical operations, i.e., linear algebra operations, the computation of the latter plays a big role in data processing. However, the current approach to deal with statistics is to export data from a DBMS to a math program, like R. At the same time the column-store approach has become popular and a number of hybrid or pure column-store systems, such as MonetDB or Apache Cassandra, are available. We want to investigate the benefits of incorporating a linear operation into a column-oriented DBMS.

The goal of this project is to integrate the QR decomposition in MonetDB, analyze the complexity of the implementation, and empirically compare the performance with the existing solutions (exporting the data to R via UDF).

The work includes the following tasks:

1. Implement a parser extension that recognizes the following new command, where R is a table name (about 1 week):
   (a) "SELECT * FROM QQR (R);"
2. Implement the part of the QR decomposition that returns table Q, using the Gram-Schmidt algorithm adapted for MonetDB system (about 3 weeks).
3. Implement table partitioning and ordering:
   - Implement a parser extension with the following command, where R is a table name, A is a list of numeric attributes of R, and O is a list of ordering attributes

(about 1 week):
"SELECT * FROM QQR (R on A order by O);"

- Implement the QQR computation on the A attributes of R, keeping the rest of the attributes in the result (about 1 week).
- Implement the sorting of R before performing the QQR computation (about 1 week).

4. Evaluate the efficiency of your implementation, i.e., evaluate the performance of the implementation and compare it to the performance of original Gram-Schmidt algorithm (about 1 week).
5. Run an experimental analysis with different table sizes (up to 100 attributes and 1'000'000 rows), varying both application and descriptive parts (about 3 weeks):
   - Analyze the runtime of the different components (ordering, trees creation, QQR execution) of your solution.
   - Implement QR decomposition using R embedded in MonetDB (R-UDF).
   - Compare the runtimes of the suggested solution and the embedded R solution.
6. Write a thesis (approximately 50 pages).
7. Present your thesis in a DBTG meeting (04.07.2017, 14:30-15:00).

Supervisor: Oksana Dolmatova
Start date: 01.03.2017
End date: 01.09.2017

University of Zürich
Department of Informatics

Prof. Dr. Michael Böhlen