Shoshin Presentation April 2014

Xu Cui

UNIVERSITY OF
WATERLOO

April 4, 2014

Agenda

- Previous Project Report
- Future Work
    - Motivation
    - Current Research Topics
    - Problems Statement
    - Sketched Approach
- Summary

Previous Project Report

- MicroFuge is near its completion.
- Spent this term polishing the paper.
- The paper is accepted to ICDCS. Special thanks to Akshay, Ben, Bernard and Khuzaima.[1]
- I shall present the paper in Madrid at the beginning of July, 2014.
- Most details of MicroFuge were presented in the previous Shoshin Meeting and I would like to talk about the work I may work on next.

[1] Ordered by alphabet

Future Work

I need your suggestions and help, feel free to speak up if you have anything to say. Anything is fine.
I am interested in all of your ideas.
Just speak up. Thanks!

Future Work - Motivation

216 systems in ranking, April 2014

Rank	Last Month		DBMS	Database Model	Score	Changes
1.		1.	Oracle	Relational DBMS	1514.08	+22.28
2.		2.	MySQL	Relational DBMS	1292.67	+2.45
3.		3.	Microsoft SQL Server	Relational DBMS	1210.43	+5.15
4.		4.	PostgreSQL	Relational DBMS	230.23	-4.82
5.		5.	MongoDB	Document store	214.34	+14.35
6.		6.	DB2	Relational DBMS	184.58	-2.74
7.		7.	Microsoft Access	Relational DBMS	142.76	-3.72
8.		8.	SQLite	Relational DBMS	90.17	-2.80
9.	↑	10.	Cassandra	Wide column store	78.72	+0.63
10.	↓	9.	Sybase ASE	Relational DBMS	78.14	-3.42
11.	↑	12.	Solr	Search engine	62.88	+1.74
12.	↓	11.	Teradata	Relational DBMS	61.73	-0.90
13.		13.	Redis	Key-value store	58.46	+5.01
14.		14.	FileMaker	Relational DBMS	54.38	+1.47
15.		15.	Informix	Relational DBMS	36.70	-0.49
16.		16.	HBase	Wide column store	36.60	+1.46
17.	↑	18.	Hive	Relational DBMS	31.02	+0.81
18.	↓	17.	Memcached	Key-value store	30.99	-1.92

Figure : Screenshot of DB Ranking from http://db-engines.com

Future Work - Motivation

- The ranking is computed based a variety of information including number of mentions on system websites, general interests based on Google trend, job offers, etc.
- NoSQL databases are gaining its popularity in today's cloud environments.
- NoSQL data storage provides native horizontal scaling.
- NoSQL data storage *provides* better performance.
- NoSQL data model addresses large volumes of structured, semi-structured or unstructured data.[2]
- etc.

Future Work - Motivation

- Query planner is one of the most important cornerstone of a RDBMS.
- Traditionally, a query planner is not needed for most NoSQL databases because there are not many tuneable knobs to be considered in the query planning besides indices.
- NoSQL databases are starting to have more tuneable knobs
  - e.g. Write Concern of MongoDB, a way to specify write consistency levels.[3]
- If we can have more optimization options, query planning should be a part of the NoSQL databases.

[3] http://docs.mongodb.org/

Future Work - Current Research Topics

- Some query optimization/planner works are done on NoSQL databases in order to combine the use of NoSQL and SQL databases in a single storage system.
  - Data Integration over NoSQL Stores Using Access Path Based Mappings, by Olivier Curé, et al., proposed a middleware system which sits on top of several NoSQL stores that processes SQL queries using Access Path Based Mappings.
  - Bridging SQL and NoSQL – A Master's Thesis by John Roijackers at Eindhoven University of Technology.
- Not much work has been done in NoSQL databases because only the very basic optimizations, nowhere as sophisticated as SQL optimization, have been done in limited number of NoSQL databases.[4]

---

[4]History Repeats Itself: Sensible and NonsenSQL Aspects of the NoSQL Hoopla by C. Mohan at IBM Almanden Research Center

Future Work - Problem Statement

Derive an utility function based on relevant cloud system states(such as network conditions), datacenter cost metrics(such as energy consumption) and SLAs(such as consistency requirement). Design and implement a query planner for the existing NoSQL databases such that it maximizes the utility function to provide the optimal query plan with respect to any prioritized view of the cost models.

Future Work - Sketched Approach

- Study the query languages of the NoSQL databases.
- Analyze all possible cost metrics.
- Derive the Utility function.
- Implement the utility function.
- Merge the utility function implementation into the query languages.
- Evaluate the performance gain by comparing the cost metrics

Summary

- An query planner for NoSQL databases.
- Only at the planning stage.
- Please provide suggestions and criticisms, the more the merrier.
- Thank you!
- *If you are willing to share with me how you look for your research ideas, please speak up. Or you can talk to me personally.*
  I would really appreciate your help. I am open to any suggestions.