Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Enterprise Platform and Integration Concepts: Research Group of Prof. Dr. Hasso Plattner

Masterproject Winterterm 2014/15

HOT or NOT?
Data Aging Re-defined.

Motivation

In-memory databases are on the rise. All major database vendors are working on in-memory database solutions. Due to decreasing main memory prices and new hardware developments, even very large systems can be entirely stored in main memory. This allows not only transactional queries to be processed in real-time, but also analytical ones – forming a so-called mixed workload, running on a single database system. However, main memory is still a comparatively scarce resource whose utilization needs to be optimized.

An immanent problem in today's in-memory databases is that data that is rarely or never accessed is handled the same way as often-requested data. As of today, mixed-workload databases have no approach for data eviction or caching. However, storing irrelevant (cold) data with the same priority as highly relevant (hot) data does not fully leverage the potential of modern server systems. If irrelevant data could be evicted to other storages based on its relevance, memory utilization and query performance can be improved.

To quantify data relevance in a real-world database system, you will have access to the workload of a large enterprise system. This productive system contains more than 100.000 tables, has a compressed-size of ~2.5 TB, and handles ~1.5 billion queries each day.

The goal of this project is to use workload characteristics in order to classify data into hot and cold partitions using so-called dynamic aging rules. By implementing a mechanism for identifying appropriate rules for a given workload, we can avoid access to cold partitions for the majority of queries. As a result, queries are answered on a fraction of the system data improving query performance, system resource consumption as well as overall memory utilization.

Project Goals

1. Familiarization with in-memory data management and hands-on experience with SAP HANA
2. Analyze data-selection characteristics of productive database workload traces
3. Implement a concept for dynamic aging rules
4. Evaluation and application of aging rules on a productive workload

Technology & Skills

Technologies and languages for the project are intentionally not restricted and will be determined during the requirements engineering. This openness with regards to technologies requires a broad set of skills. Hence, participants should have interest in any of the following:

- Python (Django, Panda, SciKit, …)
- Unix Bash/Shell Scripting
- Database Technologies (In-Memory Data Management, SQL, Stored-Procedures, ...)
- Big Data Analysis (R, SAP Predictive Analytics Library, SAP Lumira, …)

Group Structure and Project Start

The team will consist of 3-6 students. Project start will be October 13, 2014.

Contact

You are welcome to visit us in the “Villa” or reach out to one of the contacts listed below. For further information, we also invite you to an upfront meeting at room V2.16 on **July 09, 2014 at 3.15 PM**.

Dr. Matthias Uflacker (matthias.uflacker@hpi.uni-potsdam.de)
Carsten Meyer (carsten.meyer@hpi.uni-potsdam.de)
Martin Boissier (martin.boissier@hpi.uni-potsdam.de)