Research Plan

TDT39 - Empirical Research Methodology

Håkon Åmdal

December 2nd, 2015

Title of Research: Optimization Techniques for Business Discovery Products
Responsible: Håkon Åmdal, Genus AS, Svein Erik Bratsberg
Time period: Master's thesis

1 Purpose

Business Intelligence, or *Business Analytics*, allows a business to make proper decisions based on data gathered mainly in the IT systems within the company [8]. Traditionally, data warehouses have been used for this, with reports are preconfigured and generated in batches. However, with the advent of in-memory databases and cheap, commodity hardware, several market players has come up with fast, elegant, and end user intuitive solutions to analyze business data. Products like *QlikView*, *Tableau*, and *Microsoft PowerPivot* offers high-performance analytics panel which are easy to configure for the end user, such that distinctive patterns are spotted and acted upon [4, 5]. These products help people answer their stream of questions, such that they can follow their own path to business insight. We refer to these as *Business Discovery* products [5].

There are several challenges with these *Business Discovery* products. First of all, they are separate products and does not integrate well with existing solutions. Second, these applications require explicit knowledge about metadata, more specifically; types and database relations must be reconfigured when defining the data import routine. Third, to handle multiple users at once, these systems require a separate system architecture, which is yet another system for the IT staff to install and maintain. In general, the *Business Discovery* products work in isolation and are detached to the underlying data.

GENUS Application Framework is a development and architectural framework created by Genus AS for IT systems [1], and by implementing *Business Discovery* capabilities in the *GENUS Application Framework*, the above challenges can be overcome. We, therefore, work towards the following goal:

G1: Implement *Business Discovery* capabilities in *GENUS Application Framework* that have high performance, handle large datasets, and utilize available hardware. The product must be competitive to other *Business Discovery* products regarding performance and functionality.

Ongoing research (autumn project) address the following research question:

RQ1: How to design a high-performance database software that is capable of supporting *Business Discovery* workloads on large datasets?

In this research, we identify several techniques that are used to improve performance on Online Analytical Processing (OLAP) workloads. Also, we relate our discoveries to *Business Discovery* specific assumptions that are in-memory processing, read-only workloads, and special query handling. Techniques identified so far include data compression, column storage with horizontal partitioning, bitmap indexes, and multithreaded and SIMD parallelization.

For the master's thesis, we plan to follow up the autumn project by addressing the following research questions:

RQ2: What are the best optimization techniques for a *Business Discovery* product?
RQ2.1: Which optimization techniques apply to a Business Discovery product but not a general OLAP DBMS?
RQ2.2: Which optimization techniques can be combined?

Our contributions in this research are threefold. First, we integrate *Business Discovery* within *GENUS Application Framework* such that Genus AS' customers can perform *Business Discovery* within their main IT system, an integration that has not been done before. Second, we investigate and compare optimization techniques category by category, and test the effects of each optimization technique. Third, we close the gap between well studied OLAP database systems and the lesser detailed *Business Discovery* whitepapers. We look into our contributions more closely in Section 2.

2 Product

Our research contributes to an ***improved computer-based product***. As explained in Section 1, we integrate *Business Discovery* capabilities within the business' main IT system to overcome the challenges of product isolation, integration, and maintenance. We plan to integrate a system where users can access *Business Discovery* dashboards and data extracts that seamlessly interact with their main IT system. To our knowledge, such system is currently not available.

In our process, we plan to ***improve the evidence*** of optimization techniques for OLAP databases. We study techniques by category and plan to investigate the effects of each optimization technique in isolation. We know that Abadi *et al.* have studied the effects of compression, late materialization, and column store [3]. Raman *et al.* have studied the impact of SIMD processing in databases [6]. Sidlauskas *et al.* have studied the consequences of how different implementation of the same algorithm affects performance [7]. However, this research is limited in scope and does not test how the optimization techniques combine. There are also techniques that are not studied at all, as the effect of horizontal partitioning.

Lastly, we plan to ***introduce new evidence*** to optimizations that are specifically related to *Business Discovery* products. Current products do not reveal much about how they work internally. We plan on applying obtained knowledge from OLAP databases to a *Business Discovery* setting, hence adding to the body of knowledge by presenting which techniques are best for in-memory processing, read-only, and *Business Discovery* query restrictions.

3 Process

To answer **RQ2**, we combine ***design and creation*** and ***experiment*** strategies, focusing mostly on the latter. We plan to implement optimization techniques in *GENUS Application Framework* that we identified when answering **RQ1**. For each optimization technique, we ***observe*** the performance impact for each optimization in isolation, and later in combination with other techniques. We analyze our observations using a ***quantitative*** method.

We test our implementation using two test suites. The first is the standardized TPC-H benchmark that is used to test ad-hoc, decision support systems [2]. This benchmark systematically covers a broad range of industry-wide relevant queries. To answer **RQ2.1**, we perform a second test using data supplied by Genus AS' customers. Here, we develop specific test cases in collaboration with the client.

We measure performance impact in terms of query latency, query throughput, and memory footprint, and analyze our research in a quantitative fashion.

4 Participants

We execute the research with the following participants:

- Involved in this research is *Håkon Åmdal*, the main author, and department professor *Svein Erik Bratsberg*. Åmdal will be the main executor of the research and will implement, test, and document findings.
- *Genus AS* will be involved in the implementation of optimization techniques, and will provide software and hardware in which the optimizations can be implemented and tested.

- *Three of Genus AS' customers* will provide test data for the research. A consent form has been signed such that their data can be used for the purpose of this research.

5 Paradigm

We use *the scientific method*, or the *positivism paradigm*, in our research. We treat the world as ordered and regular and that it can be investigated objectively. In this research, we hypothesize the outcome of applying optimization techniques, test and observe the performance impact by measuring the execution time before and after implementation, and conclude how efficient each optimization is. We will strive to keep our research objective, repeatable, and valid.

6 Presentation

The primary deliverable in this research is a thesis documenting our findings. The thesis will include implementation details, timing results, and a brief conclusion per technique applied.

As a second deliverable, Genus AS will be left with *Business Discovery* functionality in *GENUS Application Framework* that performs well and has design decisions that are well justified.

References

[1] Genus. http://www.genus.no/?PageKey=91ff59b3-216e-4c89-8bd8-d9d366cc8479. Accessed: 2015-11-29.

[2] TPC-H - homepage. http://www.tpc.org/tpch/. Accessed: 2015-11-28.

[3] Daniel J Abadi, Samuel R Madden, and Nabil Hachem. Column-stores vs. row-stores: how different are they really? In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 967–980. ACM, 9 June 2008.

[4] Neelesh Kamkolkar, Ellie Fields, and Marc Rueter. Tableau for the enterprise: An overview for IT. Technical report, 2015.

[5] Qlik. What makes QlikView unique. Technical report, January 2014.

[6] V Raman, G Swart, Lin Qiao, F Reiss, V Dialani, D Kossmann, I Narang, and R Sidle. Constant-Time query processing. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 60–69, April 2008.

[7] Darius Šidlauskas and Christian S Jensen. Spatial joins in main memory: Implementation matters! *Proceedings VLDB Endowment*, 8(1):97–100, September 2014.

[8] Wikipedia contributors. Business intelligence. https://en.wikipedia.org/w/index.php?title=Business_intelligence&oldid=691203830, 18 November 2015. Accessed: 2015-11-22.