PROJECT PERIODIC REPORT

PUBLISHABLE SUMMARY

Grant Agreement number: 318633

Project acronym: AXLE

Project title: Advanced Analytics for Extremely Large European Databases

Periodic report: **1st x 2nd □ 3rd □ 4th □**

Period covered: from M1 November 2012 to M12 October 2013

Funding Scheme: Collaborative project

Date of latest version of Annex I against which the assessment will be made: 2012-07-16

Name, title and organisation of the scientific representative of the project's coordinator: Simon Riggs, CTO, 2ndQuadrant Limited

Tel: +44 870 766 7756

Fax: +44 870 838 1077

E-mail: simon@2ndquadrant.com

Project website: http://axleproject.eu

Publishable Summary

Rationale

The AXLE project seeks to improve Business Intelligence capabilities for Europe. AXLE's approach focuses on very large and growing databases, while addressing the full and complete requirements of real world data. Real data sources have many difficult characteristics. Data sets often grow extremely large as business initiatives succeed, so managing large data volumes once you know you have them is important.

AXLE's main concerns are

- Database Performance, Scalability & Manageability
- Security, Privacy & Auditability
- Visual Analytics & Data Mining
- Advanced Architectures for Hardware & Software

Our direction takes in novel approaches in both hardware and software that may offer a way around using brute force strategies to providing value from what has latterly become known as "Big Data".

Software features will be released as commercially-usable open source code, and submitted for wider use as core features or pluggable extensions of the PostgreSQL database (and its derivatives) and/ or Orange data mining and visualisation tool.

Validation will be carried out by industrial consortium partners with access to large volumes of private medical data, as well as standard industry benchmarks and further wide-ranging data from other interested parties. In addition, a strong Industry Advisory Board has been assembled to ensure our work has deep relevance.

Technical Expertise

The AXLE consortium includes top research and system integration organisations with non-overlapping skills in the areas of computer architecture, databases, reconfigurable systems, runtime environments, programming models and benchmark design.

The academic partners are hardware and compilation/ runtime experts. In addition, they are experts in accelerators and multi- and many-cores as well as reconfigurable computing. They provide the industry partners (who are experts in databases) with the necessary knowledge and tools to develop database engines for future architectures, as well as for the cutting edge many-core processors of today.

Intended Target Groups and Domain

AXLE targets databases which contain **Important** data, and thus will be **Complex**, which when successfully used will become **Extremely Large**, which will in turn require strong **Privacy & Security** controls.

The improvements will focus on functionality and performance for use in business intelligence applications on very large datastores, especially with the proviso that transforming and re-formatting data into a data warehouse is not a viable option at very large data volumes.

Expected Results

The AXLE project aims to greatly improve the speed and quality of decision making on real-world data sets and to make those improvements generally available through high quality open source implementations via the PostgreSQL and Orange products.

AXLE will deliver:

- Advanced analytical hardware/software techniques that show significant measurable improvements in database processing speed over existing techniques when applied to extremely large and realistic data volumes.
- Advanced techniques for addressing the scalability challenge of extremely large datasets, specifically the ability for many common queries to return in the same time no matter how large the data by using flexible proof-based approaches to query handling.
- Visual analytics capable of exploring extremely large data volumes without significant loss of speed or functionality as data volumes grow.
- A capability to measure and evaluate performance against extremely large volumes of real data, with a mechanism for more easily publishing and comparing results.
- More scalable data management with integrated security controls.
- High security database software capable of securing and auditing data in its application context, as well as pass external assessment as being suitable for Common Criteria for IT Security Evaluation. http://www.commoncriteriaportal.org/

Results so far

- Benchmark environments fully set-up
  - Benchmark workloads and data sets specified in detail, including appropriate anonymisation to ensure privacy of real data
  - Hardware resources purchased and usable
  - Benchmark recording and analysis application available
  - Baseline benchmarks suggest that sorting, hash joins, fixed point arithmetic operations, compression/decompression and deform tuple operations could profit from hardware acceleration
- Orange is now working with PostgreSQL
  - A basic interface between Orange and PostgreSQL has been implemented ahead of schedule.
  - A radical new approach to data mining transformations allows direct generation of SQL, moving the workload completely into the database
- Production ready implementations have been submitted to PostgreSQL 9.4
  - MinMax Indexes - a new index type towards automatic partitioning for very large and growing database tables
  - Event Triggers for auditing of data definition statements
  - Scalability improvements for core PostgreSQL locks and memory
- Initial prototypes are working towards production readiness for
  - Row Level Security
  - On-disk Bit Map Indexes
- Preliminary analysis suggests that machine learning could be applied at different levels to improve the performance of the database. Selecting an appropriate algorithm/ technique for the different functionalities (currently sorting) or foreseeing accesses (currently the next memory address) are just a few examples of where machine learning can have a significant impact.
- A prototype sorting algorithm that exploits JIT vectorization to boost performance by 2X on the average has been implemented.
- Sorting implementations on FPGA have been completed. Working on a paper that compares the different hardware design language approaches, that will be sent to the FCCM 2014 conference.

At a glance

Project title: AXLE - Advanced Analytics for EXtremely Large European Databases
Project coordinator: Simon Riggs, 2ndQuadrant Limited (UK)
Partners: 2ndQuadrant Limited (UK), Barcelona Supercomputing Center (ES), Portavita B.V. (NL), The Univ. of Manchester (UK), The Univ. of Ljubljana (SI)
Duration: 1st November 2012 - 30 October 2015
EU contribution: EUR2.9M
Further information: http://axleproject.eu