AXLE
Advanced Analytics for Extremely Large European Databases
axleproject.eu

Type of the presentation proposed:
Research contribution

Title of the presentation:
Big Data - can it be tamed?
(Automatic scaling of complex analytics for real-world data)

Summary of the presentation (<100 words)
Analytics on European level datasets is critical to competitive industry. Minimising costs to society of government, medical, science and military endeavours requires pooling of data for analysis. Larger datasets require faster analytical processing and demand cost effective solutions.

The AXLE Project (Advanced Analytics on Extremely Large European Databases) explored novel approaches in 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”.

This presentation will outline the enhancements made and the results of the AXLE Project.

Extended abstract of the presentation

AXLE focuses on automatic scaling of complex analytics, while addressing the full requirements of real datasets. Real data sources have many difficult characteristics. Sources often start small and can grow extremely large as business/ initiatives succeed, so the ability to grow seamlessly and automatically is at least as important as managing large data volumes once you know you have them. Extremely large data stores have added concerns such as data quality, privacy, security and auditability.

Aspects of the project include
* Scalability Engineering - Autopartitioning, Compression
* Security, Privacy and Audit Techniques
* Visual Analytics
* Advanced Architectures for Hardware and Software

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

The overall strategic impact we aim for is to produce a viable open source alternative to commercial solutions for Big Data, with a product set based around the PostgreSQL enterprise database.

AXLE brings together a diverse group of researchers covering hardware, database kernel and visualisation experts, all focused on solving the needs of extremely large data analysis. The 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 (University of Manchester, Barcelona Supercomputing Center and University of Ljubljana) are hardware and compilation/ runtime experts. In addition, they are experts in accelerators and multi- and many-cores as well as reconfigurable computing. Therefore, they are in a position to provide the industry partners (who are experts in databases - 2ndQuadrant and Portavita) with the necessary knowledge tools to develop database engines for future architectures, as well as for cutting edge multi- and many-core processors of today.

The AXLE project is at the forefront of addressing the challenges of new architectures, illustrating how hybrid hardware architectures can be used in a cloud/ virtual environment to resolve performance issues. This shows the way forward for public or large private infrastructures where small numbers of specialised hardware appliances can make a huge difference to normal workloads, without the need for widespread purchases of new or enhanced hardware.

Reduced cost and practical increases in capability to manage large data volumes will enable new scientific investigations, both in academia and in commercial business analysis or R&D. The improvements delivered by AXLE will allow faster responses to database queries allowing more timely decision making. Decision-making will be improved by

allowing more questions to be asked by a wider pool of users. Practical direct access to detail data will reduce the need for copies, extracts and data samples which inevitably lead to mistakes and reduce effectiveness of data based decisions. Lastly, visual analytics against large data volumes will make conclusions more obvious and allow faster group agreement on insights and the need for action, leading to more timely and better decision-making.

A broad outline of the presentation is below:

- Introduction covering the methodology followed in the project
- Short video (or live demonstration) of how the optimised process works
    - illustration of the benefits of academia and industry collaborating
- Project results
- Conclusion
- Q&A session with audience if appropriate

The main outcomes of the presentation will be:

- A shared understanding of the challenges and opportunities for Big Data analytics in Europe
- Dissemination of the AXLE Project's results

The targeted audience will be:

- Data analysts from research, academia and industry
- Industry leaders
- infrastructure and network managers
- Senior IT budget holders in medium and large enterprises

Contributor names and short CVs

The AXLE Consortium is comprised 5 partners organisations, the details of which are listed here. The short cv of the lead/s in each partner is provided, although there are further contributors from each organisation (details can be provided if required).

2ndQuadrant

2ndQuadrant is the largest worldwide team of contributors to the PostgreSQL open source database project. 2ndQuadrant is headquartered in Oxford, UK, with offices across Europe and worldwide.

2ndQuadrant is a professional services company and software house, specialising in large enterprise-class database issues, particularly PostgreSQL and derivative products. 2ndQuadrant is a key sponsor and development contributor of PostgreSQL, and has been for more than 10 years. 2ndQuadrant staff are both practical solutions implementors and researchers into the features needed by the largest users of database technology for complex solutions. Specific research interests are database performance, high availability and scalability.

Simon Riggs is Chief Technology Officer (CTO) and founder of 2ndQuadrant. Having worked as a database architect for more than 25 years in industry, Simon is a Major Developer of the PostgreSQL open source database project. He is one of the few committers with authority to commit software changes to the PostgreSQL core code, acting as maintainer for the recovery and replication features.

Simon personally contributed major features in each of the last 10 annual releases of PostgreSQL.

Simon is also lead author of "PostgreSQL 9 Admin Cookbook", Packt Press. Previously, Simon worked for Teradata and was the architect of the first data warehouse at British Airways. Simon's early research work in Astrophysics was published by the Royal Society.

Alvaro Herrera started improving the PostgreSQL code base in 2002, producing many features in every release since and continues to be a significant member of the core development team. He is also a PostgreSQL committer and wrote a significant PostgreSQL feature, BRIN Indexes, for the AXLE Project.

BSC

The Barcelona Supercomputing Center, established in 2005, serves as the National Supercomputing Facility in Spain. The mission of BSC is to research, develop and manage information technologies in order to facilitate scientific progress. The Computer Sciences Department of the BSC focuses on building upon currently available hardware and software technologies and adapting them to make efficient use of computing infrastructures. The department proposes novel architectures for processors and memory hierarchy, develops programming models and innovative implementation approaches for these models, and develops tools for performance analysis and prediction.

The Computer Architecture for Parallel Paradigms research group specializes in computer architecture and hardware in general, with special applicability to designing multi-core systems for performance and low- power while leveraging novel programming models and runtimes.

Dr. Osman Unsal is co-manager of the Computer Architecture for Parallel Paradigms research group at BSC. He received the B.S., M.S. and PhD degrees in Electrical and Computer Engineering from Istanbul Technical University (Turkey), Brown University (USA) and University of Massachusetts, Amherst (USA) respectively. Together with Dr. Adrian Cristal, he co-manages the Computer Architecture for Parallel Paradigms research group at BSC. His current research interests include many-core computer architecture, reliability, low-power computing, programming models and transactional memory.

Dr. Adrián Cristal is co-manager of the Computer Architecture for Parallel Paradigms research group at BSC. His interests include high-performance microarchitecture, multi- and many-core chip multiprocessors, transactional memory, and programming models. He received a PhD from the Computer Architecture Department at the Polytechnic University of Catalonia (UPC), Spain, and he has a BS and an MS in computer science from the University of Buenos Aires, Argentina.

Dr. Adrià Armejach is a post-doctoral researcher at the BSC. His interests include computer architecture, parallel computing, memory systems, and performance evaluation. He received a PhD from the Universitat Politècnica de Catalunya (UPC) in 2014 and is currently involved in providing solutions using emerging memory technologies in the context of an FP7 project on Advanced Analytics for Extremely Large European Databases (AXLE).

Dr. Nehir Sonmez holds a PhD on Computer Engineering (2012) from the Technical University of Catalonia (UPC), an MS degree from Bogazici University (2006) and a BS degree from Syracuse University (2003). He is currently a post-doctoral researcher in the Computer Architecture for Parallel Paradigms research group at the Barcelona Supercomputing Center. His research interests include reconfigurable computing, computer architecture and multicore, database acceleration for big data and transactional memory.

Portavita

Portavita is based in Amsterdam, Netherlands and is fully focused on the development, sales and implementation of its web-based Disease Management System, DMS, which supports multidisciplinary, integral treatment of people with chronic conditions. Based on input from leading physicians and their patients, Portavita develops the various modules of the DMS, including diabetes, asthma/COPD, cardiovascular risk management and anticoagulation. About 198,000 active patients are being treated with the help of Portavita DMS as of January 2012. Portavita is the market leader in this area.

Good cooperation between all those providing care for the (chronic) patient is essential if the care challenges are to be met. The Disease Management Systems (DMS) follows best medical practices, treatment protocols and openly developed standards such as HL7, SNOMED CT and LOINC. Portavita is making significant contributions to the open standardization process in the care sector through HL7 and NEN. Portavita is also driving

IT standardization in the care sector, such as the 'HIS/KIS convenant' (GP and chain-wide information systems) and the Detailed Clinical Models project.

Evert Jan Hoijtink is Founder and CEO of Portavita BV and Founder of medical database technology company MGRID BV. He is also founder and shareholder of E3Ventures BV, a venture capital organization investing in innovative eHealth companies. In his role as Vice-Chairman of the Dutch Branch organization for IT Suppliers in the Health industry, Evert Jan
is responsible for the portfolio Standardization. He is member of the Policy Committee BC303 / HealthIT of the NEN, the Dutch Chapter of ISO. Current standardization focus of OIZ is representing the industry position paper in the Dutch EHR discussions and an initiative on Semantic Interoperability for medical observations in cooperation with NICTIZ the National IT Institute for Healthcare in the Netherlands, ZN the Dutch Health Insurance Branch Organization (www.zn.nl), KNMG the Royal Dutch Medical Association and NPCF the Dutch patient organization. He was trained in Economics at the Dutch school of Higher Business Education. After a function as Industrial Accountant at Akzo, he worked mainly in the IT industry as sales at Honeywell-Bull, account manager and senior consultant at CapGemini and director of eCommerce at Oracle EMEA. His core competence is business development, IT Architecture, Standardization and being an entrepreneur.

Yeb Havinga has been the software architect and technical lead of Portavita's Disease Management System, and is currently developing Portavita's next generation clinical database platform called MGRID. Yeb is also an active member of the RIMBAA Working Group of the HL7 standards organization, that focuses on application and database design based on the HL7v3 Reference Information Model (RIM) standard. Before joining Portavita, Yeb co-founded DD&H, where he was the technical director and pioneered database-driven
websites. His professional software engineering career started at Mediacenter Automatisering, where he developed for IBM's AS/400 database system. Yeb brings almost 20 years of database and software engineering expertise, with a focus on formal methods and medical informatics. Yeb holds a Master's degree in computer science and has also completed several courses from the MSc in Logic and MSc in Medical Informatics at the University of Amsterdam.

Olivier Marchesini is Quality & Compliance manager at Portavita. He joined the company in
April 2010 and is responsible for quality development and compliance to the ICT and Health
norms and standards. He participates in the NEN (Dutch CEN correspondent) committees on ICT in health care and Software as medical devices. He worked previously has consultant
at Philips, Cap Gemini and A.T. Kearney and as manager at a health insurance company (Agis Verzekeringen) and a health knowledge institute (Prismant) Olivier has a degree in mechanics from the Ecole Centrale de Lyon (FR) and in business administration from the University of Twente (NL).

UNIMAN

The University of Manchester is one of the top research-led universities and can lay claim to 25 Nobel Prize winners amongst its current and former staff and students, including 4 current Nobel laureates. The School of Computer Science plays important roles in the two

EU FET flagship projects (Graphene and Human Brain Project) and collaborates with the Square Kilometer Array (SKA) experiment headquartered in the university's Jodrell Bank Observatory. The school also has a long and distinguished research record, including the development of the first stored program computer the late '40s, and the development of virtual memory among a range of innovations in the Atlas computer in the early '60s (the UK first supercomputer).

The school retains strong activities in computer systems and engineering (indeed, graphene, the discovery of which led to the Nobel Prize for Physics in 2010, was first observed using a microscope in our engineering and nanotechnology labs). The Advanced Processor Technologies group (APT) continues the excellent record in high performance low-power computer systems, and encompasses a range of research activities addressing the formidable complexity of the many-core systems of the future. The APT group brings together more than 60 researchers (faculty, fellows, PhD students) and is one of the few centers of excellence able to design complex silicon as demonstrated by SpiNNaker; a one million ARM cores massively parallel architecture. APT has helped the EU competitive position with commercialization examples such as the ICL Goldrush Database server, Amulet processors (Low-power architectures) bought by ARM Ltd., Transitive Corporation (Virtualization and Binary Translation) bought by IBM and Silistix Ltd (Networks-on-Chip).

Dr. Mikel Lujan holds a prestigious Royal Society University Research Fellowship in APT investigating low power many-core systems. He leads the UK-funded DOME project investigating fault tolerant many-core systems and is co-Investigator in the UK-funded PAMELA project (5-year grant). At the European level, he leads in Manchester the EU STREP FP7 project AXLE investigating acceleration of analytics for large European databases. He also contributes to the EU research community as part of the RETHINK big CSA FP7 project, which is generating the EU roadmap for hardware/software codesign for Data Intensive Applications. Previously Mikel worked on run-time systems for HPC peta-scale systems for Sun Microsystems Research Laboratories (CA, USA). This work generated three US patents and was funded by DARPA/DOE. Since his first paper in OOPSLA 2000, Dr. Lujan has authored more than 60 refereed papers. His last PhD student received the UK 2013 Best PhD Thesis in Computer Science award.

Dr. Javier Navaridas is a Lecturer in computer architecture in the Advanced Processors Technologies group at the University of Manchester. He obtained his PhD in Computer Engineering in 2009 from the University of the Basque Country, which was rewarded with an Extraordinary Doctorate Award. He joined the University of Manchester with a prestigious Newton Fellowship in 2010. His main interest are improving the efficiency of interconnection networks in the context of large-scale computing systems, emerging technologies for on-chip interconnects, enabling Big Data analytics by different architectural approaches, modeling systems and workloads and application scheduling and mapping. Dr. Navaridas is leading the technical work of the INPUT EPSRC-funded project, which deals with datacentre networks and is co-investigator in AXLE FP7 project. He is also involved in RETHINK.big EU initiative.

UL

The University of Ljubljana is the oldest and largest in Slovenia, and is among the largest universities in Europe. The Laboratory of Bioinformatics employs 17 professors, researchers and software developers, and performs research in computational methods for data analysis, data mining, visualization, and artificial intelligence. Laboratory members

apply these techniques to current problems in molecular and systems biology, in particular in functional genomics and mutant-based analysis (with Baylor College of Medicine, Houston, USA), sequence and gene expression regulation analysis (with MRC Laboratory of Molecular Biology, Cambridge, UK), and tissue engineering (with University of Pavia, Italy).

The Laboratory develops and maintains a suite of software tools: the comprehensive Python-based Orange data-mining suite (http://orange.biolab.si), the epistasis analysis toolbox GenePath (http://www.genepath. org), and the gene expression web-analytics application dictyExpress (http://www.ailab.si/dictyexpress). The Laboratory has published in the highest-ranked journals in the field, including Nature (Genetics, Neuroscience, Structural & Molecular Biology), Bioinformatics, Journal of Machine Learning Research, EIII Transactions on Pattern Analysis and Machine Inteligence and Nucleic Acids Research. The Laboratory is primarily founded from grants by the Slovene Research Agency, European Grants (CARE-MI FM7 project, subcontractor to one ESF grant), and from private funding (subcontractor to Baylor College of Medicine; two grants from AstraZeneca).

Dr. Janez Demšar is Assistant Professor at the Faculty of Computer and Information Science and a member of the Bioinformatics Laboratory. He teaches courses on data mining, project management, and programming. He is the creator and Chief Architect of the successful Orange data-mining and visualisation suite and remains involved in the project as the principal contributor to the system's core. He has (co)authored over 30 peer-reviewed journal and 50 conference papers. His current research interests include intelligent visualization techniques, qualitative modelling, and modelling of high-dimensional data.

Dr. Blaž Zupan, PhD is a Professor at UL, is the head of the Bioinformatics Laboratory and serves as a Vice Dean for Research (2006-2008, 2010-2012). From 2010, he is also a Visiting Assistant Professor at Department of Molecular and Human Biology at the Baylor College of Medicine, Houston, USA; he was also awarded a Fulbright Scholarship in 2013/14. His principal expertise is in machine learning, data visualization, and data mining. He has (co)authored over 70 peer-reviewed journal and 70 conference papers, and was a PI for the Slovenian part of the group in several EU projects.