Big Data @ VLDB 2012

Tilmann Rabl

msrg.org

tilmann.rabl@utoronto.ca

Overview

- Some statistics
- Big Data Applications
- Big Data Systems
- Experimental Setups
- Workshops

Big Data in Numbers

- VLDB Big Data sessions
  - 2 industry, 1 demo, 2 tutorial, 1 workshop
- “Big Data” in paper titles/abstracts
  - 13 / 192 = 7 %

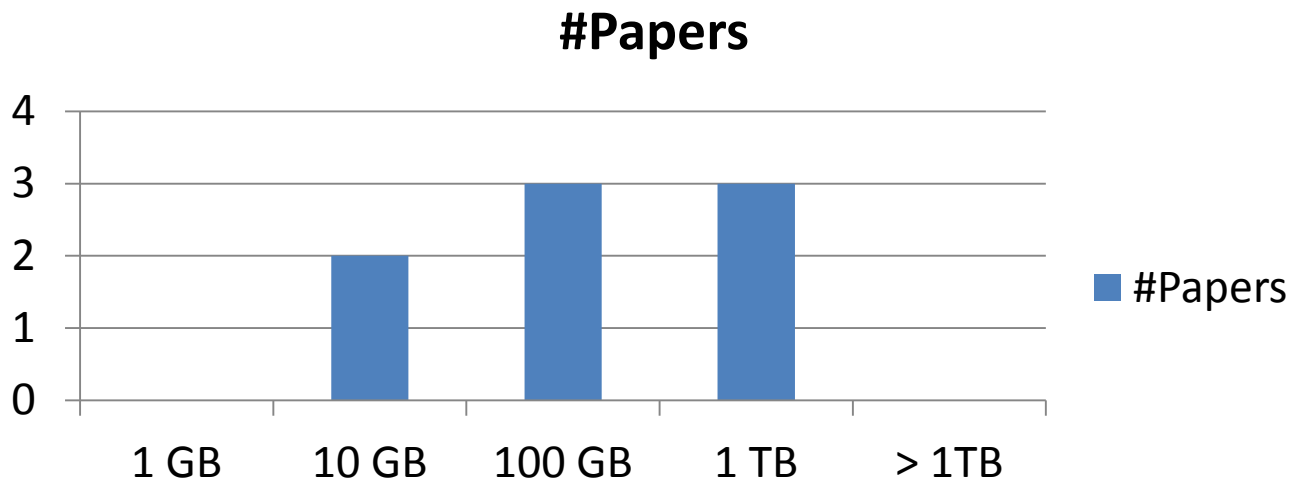
Big Data Research

- Map Reduce
- Key-Value Store
- Parallel Database System
- Query Languages
- Scientific Databases
- Energy Efficiency
- System Architectures

Big Data Applications

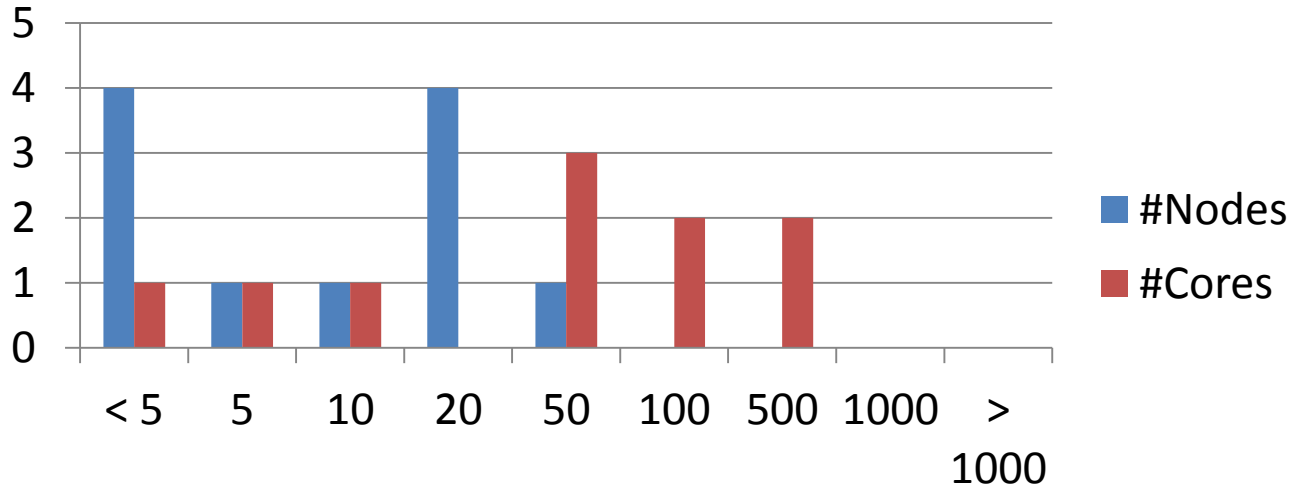
- Monitoring
  - Application performance management
  - Smart meter data
- Spatial Data
  - Satellite images
- Time Series
  - Stock prices
- Text Processing
  - News
- Graph Data
  - Link data (e.g. Wikipedia)
- Matrices
- TPC Benchmarks
  - Predominantly TPC-H

Experimental Setups I



- Maximum data sizes in experimental evaluation
- Total number of papers 8

Experimental Setups II



- Maximum number of nodes / cores in the experiments
- Total number of papers 11

Big Data Panel

- DB community missed the trend
- Big data is a hype coined by non-db people
- Big data has no real new challenges
- Big data is a buzz-word (Big is the new very large)
- Big data is different because of fault-tolerance et al.

Mike Carey Keynote TPCTC I

- Fair Tuning is Critical
- Expect Unhappy Developers
- Just How Declarative is a Query?
- Is This a Reasonable Data Set?
- Steady As She Goes!
- Single- *vs. Multi-User Performance*
- Should the World Be Open and/or Classless?

Mike Carey Keynote TPCTC II

Future BDMS Benchmark Thoughts

- Richer BDMS micro-benchmarks
  - More comprehensive than CALDA, including a broader range of queries (small/medium/large), also updates, and also indexing
- Multi-user BDMS benchmarks
  - Profile shared operational clusters' workloads to understand and evaluate real mixes and scheduling successes
- Domain-centric Big Data benchmarks
  - Test large graph analytics, scientific data, spatial data, streaming data (*a.k.a.* "Fast" or "high-velocity" data), ML tasks, etc.
- Self-management benchmarks
  - Auto-management of storage as data and/or nodes arrive, as well as during failure scenarios
- Challenging data benchmarks
  - Flexible schema support, fuzzy searching/matching, spatial data handling, etc., as needed for "SoLoMo" and other emerging trends

BigData 2012

- **Towards End-to-End Data Analysis in GeoSciences**, Tanu Malik, *University of Chicago, USA*, Neil Best, *University of Chicago, USA*, Ian Foster, *University of Chicago, USA*
- **Incremental Loading: Access-Driven Data Transfer from Raw Files into Database System,** Azza Abouzied, *Yale University, USA,* Daniel Abadi, *Yale University, USA*, Avi Silberschatz, *Yale University, USA*
- **Jet: An Embedded DSL for High Performance Big Data Processing**, Stefan Ackermann, *ETH Zurich, Switzerland*, Vojin Jovanovic, *EPFL, Switzerland*, Tiark Rompf, *EPFL, Switzerland*, Martin Odersky, *EPFL, Switzerland*
- **Massive Smart Meter Data Storage and Processing on top of Hadoop**, Leeley D. P. dos Santos, *EDF R&D, France*, Alzennyr G. da Silva, *EDF R&D, France*, Bruno Jacquin, *EDF R&D, France*, Marie-Luce Picard, *EDF R&D, France*, David Worms, *Adaltas, France*, Charles Bernard, *EDF R&D, France*
- **On Aligning Massive Time-Series Data in Splash**, Peter J. Haas, *IBM Research - Almaden, USA*, Yannis Sismanis, *IBM Research - Almaden, USA*
- **Meteor/Sopremo: An Extensible Query Language and Operator Model**, Arvid Heise, *Hasso Plattner Institute Potsdam, Germany*, Astrid Rheinlaὶˆnder, *Humboldt-Universitat zu Berlin, Germany*, Marcus Leich, *Technische Universitat Berlin, Germany*, Ulf Leser, *Humboldt-Universitat zu Berlin, Germany*, Felix Naumann, *Hasso Plattner Institute Potsdam, Germany*

Big Data Papers I

- Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads *Yanpei Chen, Sara Alspaugh, Randy Katz*
- Solving Big Data Challenges for Enterprise Application Performance Management *Tilmann Rabl, Mohammad Sadoghi, Hans-Arno Jacobsen, Sergio Gómez-Villamor, Victor Muntés-Mulero, Serge Mankowskii*
- CloudVista: Interactive and Economical Visual Cluster Analysis for Big Data in the Cloud *Huiqi Xu, Zhen Li, Shumin Guo, Keke Chen*
- ASTERIX: An Open Source System for "Big Data" Management and Analysis *Sattam Alsubaiee, Yasser Altowim, Hotham Altwaijry, Alexander Behm, Vinayak Borkar, Yingyi Bu, Michael Carey, Raman Grover, Zachary Heilbron, Young-Seok Kim, Chen Li, Nicola Onose, Pouria Pirzadeh, Rares Vernica, Jian Wen*
- Early Accurate Results for Advanced Analytics on MapReduce *Nikolay Laptev (University of California, Los Angeles, USA), Kai Zeng (University of California, Los Angeles, USA), Carlo Zaniolo (University of California, Los Angeles, USA)*
- Spinning Fast Iterative Data Flows *Stephan Ewen (Technische Universität Berlin, Germany), Kostas Tzoumas (Technische Universität Berlin, Germany), Moritz Kaufmann (Technische Universität Berlin, Germany), Volker Markl (Technische Universität Berlin, Germany)*
- Optimizing I/O for Big Array Analytics *Yi Zhang (Duke University, USA), Jun Yang (Duke University, USA)*

Big Data Papers II

- Opening the Black Boxes in Data Flow Optimization *Fabian Hueske (Technische Universität Berlin, Germany), Mathias Peters (Humboldt-Universität zu Berlin, Germany), Matthias Sax (Humboldt-Universität zu Berlin, Germany), Astrid Rheinländer (Humboldt-Universität zu Berlin, Germany), Rico Bergmann (Humboldt-Universität zu Berlin, Germany), Aljoscha Krettek (Technische Universität Berlin, Germany), Kostas Tzoumas (Technische Universität Berlin, Germany)*
- Towards Energy-Efficient Database Cluster Design *Willis Lang (University of Wisconsin, USA), Stavros Harizopoulos (Nou Data, USA), Jignesh M. Patel (University of Wisconsin, USA), Mehul A. Shah (Nou Data, USA), Dimitris Tsirogiannis (Microsoft Corporation, USA)*
- Can the Elephants Handle the NoSQL Onslaught? *Avrilia Floratou (University of Wisconsin - Madison, USA), Nikhil Teletia (Microsoft Jim Gray Systems Laboratory, USA), David J. DeWitt (Microsoft Jim Gray Systems Laboratory, USA), Jignesh M. Patel (University of Wisconsin - Madison, USA), Donghui Zhang (Paradigm4, USA)*
- Muppet: MapReduce-Style Processing of Fast Data *Wang Lam (@WalmartLabs, USA), Lu Liu (@WalmartLabs, USA), STS Prasad (@WalmartLabs, USA), Anand Rajaraman (@WalmartLabs, USA), Zoheb Vacheri (@WalmartLabs, USA), AnHai Doan (University of Wisconsin-Madison, USA)*
- MonetDB/DataCell: Online Analytics in a Streaming Column-Store *Erietta Liarou (Centrum Wiskunde & Informatica, Netherlands), Stratos Idreos (Centrum Wiskunde & Informatica, Netherlands), Stefan Manegold (Centrum Wiskunde & Informatica, Netherlands), Martin Kersten (Centrum Wiskunde & Informatica, Netherlands)*