

DIR & XLDB

Data Intensive Research eXtremely Large DataBases

Jens Jensen, NGS
NGS Surgery
15.06.2011

See also Aaron's summary, he attend the 1st day of DIR

ADMIRE project - Malcolm

- Workflow, processes: DISPEL language
 - Need representation language to be able to discuss work – hence DISPEL
- Support hot data
- MyExperiment tracks [workflow] provenance and attribution

Data Talks

- Alex Szalay – Johns Hopkins
 - Tiered data centres, user driven
 - 100 TB “hard,” 300 TB “impossible”
- Alistair McGowan (Glasgow)
 - How to convince funding bodies data is worth doing
- Michael Wise (LOFAR)
 - Raw 13 Tb/s, or 138 PB/d 😊 (2.5-3 PB/yr)
 - HDF5 “sky cubes”
 - “Transient” database: MonetDB, event processing

Database Talks

- Dave Pearson (Oracle)
 - “Sharding” – processing chunks
 - Columnar data compression – “10x for active, 15-50x for archive”
- Complex event processing – streams, pattern matching. In-memory continuous queries.
- New breed of DPMS
 - Cassandra, MongoDB, Riak, Hadoop, CouchDB
 - Schemas appearing

Database Talks

- Stuart Owen, Manchester (Taverna, MyGrid)
 - Sysmo Seek from Sysmo consortium (systems biology)
 - sysmo-db.org, www.sysmo.net
 - Rightfield for Excel ontology, www.rightfield.org.uk
 - Hidden and super-hidden sheets(!)
- Tamas Budavari, Johns Hopkins
 - Doing statistics on large datasets: R within Postgres

Background

- Data Intensive Research
 - ESI theme, closing workshop
- XLDB
 - eXtremely Large Databases (~PB)
 - All database flavours, not just relational
 - Both industry and academia
 - Database vendors (Oracle, MonetDB, ...)
 - Users: NCAR,
 - Data people, dataset owners

Database Flavours

- Database flavours
 - Relational
 - Array, object
 - Shared nothing (open source implementation)
- Subject of lively debate in db community
- “What really works”
- What will we use (on new projects)? How do we choose?

Database Use @ STFC

- Oracle
 - 3.7 TB over 17 databases
 - 80% is T1
 - 284 txns/s, ~1000 concurrent connections
- Datastore
 - ~7 PB on tape, nominal capacity of ~50-100 PB
 - ~3 PB on disk, capacity of ~8 PB

Selected Key Points

- “Mobilising resources is key, not funding – but funding is a good way of mobilising resources”
- Datasets, sharing
 - How to credit essential work on datasets
 - Provenance, attribution
 - “Social” and “cultural” barriers, legal
 - Quality
 - Multitude of data sharing initiatives
- Exploit existing infrastructures
- Open source, standards. Hackathon.

Reference

- DIR theme:
 - wiki.esi.ac.uk/Data-Intensive-Research_Theme
- Agenda:
 - www.xldb.eu/xldb_europe_2011/program.html