Scientific Big Data Management Research

Tore Risch

Uppsala DataBase Laboratory (UDBL)
(http://www.it.uu.se/research/group/udbl/)
Department of Information Technology
Uppsala University, Sweden

The Big Data flood

Enormous data growth

The traditional Moore's law:
- Processor speed doubles every 1.5 years

Current data growth rate *much* higher
- Data grows 10-fold every year!

Enormous data growth

Major *opportunities:*
- spot business trends
- prevent diseases
- combat crime
- scientific discoveries, the *4th paradigm* *http://research.microsoft.com/en-us/collaboration/fourthparadigm/*
- data-centered economy

Major *challenges:*
- Information overload
- *Scalable* data processing, 'Bigdata management'

See: Economist 2010-02-27:
http://www.economist.com/node/15557443

UDBL, Uppsala DataBase Laboratory

Mission:

Development of *software* and *algorithms*

to enable *search* and *analysis* of *big data* volumes

in *distributed* and *heterogeneous* environments.

Application scenario I

- High volume streams of many measurements received
  - E.g. patient monitoring
- Expensive numerical models over streams to predict abnormal behaviors on-the-fly
  - E.g. models to predict high likeliness of stroke or heart attacks within the next day
- Continuously run models over lots of data
  - E.g. CQs over all patients in hospital, city, country,…

Data Base Management Systems

SQL Queries

DBMS

Query Processor

Data Manager

Meta – data

Stored Data

Data Stream Management Systems

Continuous Queries (CQs)

DSMS

Query Processor

Data & Stream Manager

Data streams

Data streams

Meta – data

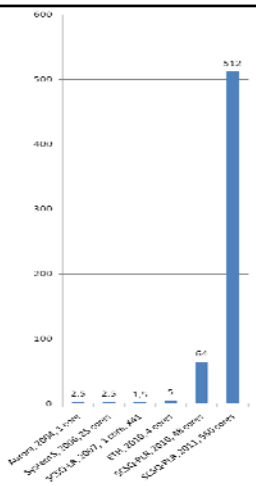
Stored Data

Performance

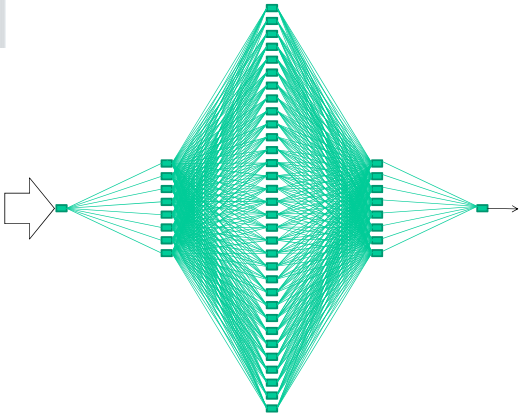
Linear Road

A Stream Data Management Benchmark

www.cs.brandeis.edu/~linearroad



How: Massive parallel stream processing



Federated DSMS engine (FDSMS)

Programming languages: Java Python C LabView Matlab

Query languages: SPARQL SQL SCSQL

External languages

Algorithm plug-ins

FDSMS:

Local Semantic DB

(Continuous) Query API

FDSMS kernel

Plug-in manager

Python Java Lisp C

Indexing

Prediction

Matching

Optimization

Classification

Wrapper manager

Stream wrappers Data wrappers

Data sources: JSON feed FDSMS JDBC NetCDF

Federated DSMS engine (FDSMS)

Programming languages: Java Python C LabView Matlab

Query languages: SPARQL SQL SCSQL

External languages

Algorithm plug-ins

FDSMS:

Local Semantic DB

(Continuous) Query API

Stream Validator

Query Processor

Distribution engine

Query plan interpreter

Plug-in manager

Python Java Lisp C

Indexing

Prediction

Matching

Optimization

Classification

Wrapper manager

Stream wrappers Data wrappers

Data sources: JSON feed FDSMS JDBC NetCDF

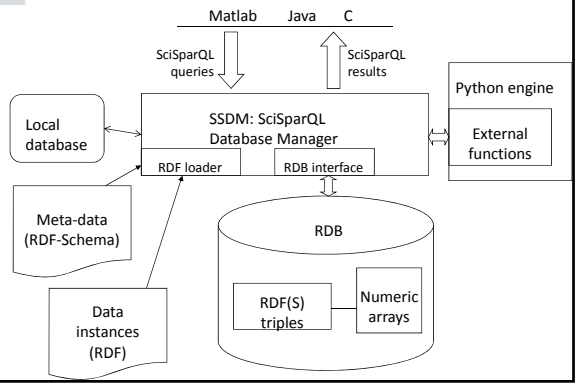
Application scenario II

- Large volumes of scientific experiments producing large volume numerical (e.g. matlab) data
  - E.g. trajectories, matrices
- Meta-data describing properties of experiments
  - E.g. subjects, units used, description, steps
- Queries combining meta-data with experimental data
  - E.g. trajectories crossing through an area for some specific kind of experiment

SciSPARQL Data Manager (Andrej Andrejev)

•The query language SPARQL is standard semantic web query language
  - From CERN for science
  - Very well suited to describe (e.g. experimental) **meta-data**
  - Not well suited for querying numerical data
- Defined and implemented extended query language SciSPARQL
  - Very strong on querying numerical data
  - Allows querying and analyzing *combined* meta-data and experimental data
- SciSPARQL implemented in SSDM

Scientific SPARQL Data Manager (Andrej Andrejev)



Application scenario III

- Scientific experimental data and meta-data stored in relational database
  - E.g. micro data, environmental data, economical data, etc.
- Preserve *selected results* of experiments for long term in neutral format
  - E.g. using the RDF Schema semantic web standard meta-data format
- Restore database in other future formats
  - E.g relational databases, RDF stores

Semantic Archive and Query system (Silvia Stefanova)

SAQ:

•Long term preservation of scientific data in relational databases

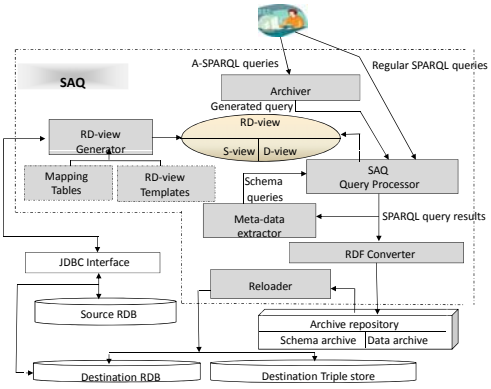
•Preserve using semantic web data model RDF

•Selective preservation of both *data* and *meta-data* using the extended A-SPARQL query language

•Preserved data and meta-data can be later restored in future database

•Provides migration path between traditional relational DBs and e.g. RDF stores

Semantic Archive and Query system (Silvia Stefanova)



Thank you
For your attention

Tore Risch

http://user.it.uu.se/~torer