LARGE-SCALE DATA ANALYTICS
Applications and Technology

Peter Brezany

Research Group Scientific Computing
Faculty of Computer Science
University of Vienna, Austria

2nd SPLab Workshop, October 2012
Today’s Data Flood

- Business
- Medicine
- Earth Observations
- Scientific Experiments

- Data Mining
- Selection
- Processing
- Integration
- Simulations
Big Data

• Every day, we create 2.5 quintillion ($10^{18}$) bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data is big data. The associated analytic processes are called large-scale data analytics.

• Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, manage, and process the data within a tolerable elapsed time.

• Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes ($10^{12}$) to many petabytes ($10^{15}$) of data in a single data set.
Towards High-Productive Analytics

• This term firstly appeared in the context of e-Science analytics.

• **E-Science Analytics** is a dynamic research field that includes rigorous and scientific methods of data preprocessing, integration, analysis, and visualization.

• A **high-productive analytics system** is one that delivers a high level of performance, guarantees a high level of accuracy of analytics models and other results extracted from analyzed data sets while scoring equally on other aspects, like usability, robustness, system management, and ease of programming.
Response: Data-Intensive Research

• Offers powerful methods aiming at innovative new uses of computing, storage and network devices when capturing, managing, analyzing, and understanding data produced by modern science and business at rates that push the frontiers of current technologies.

• There is a number of research challenges.

Grid- and Cloud-Based Data Analytics

• **GridMiner** ([www.gridminer.org](http://www.gridminer.org))
  – is based on the Globus 4 Toolkit, OGSA-DAI.
  – includes: BPEL-based workflow engine, distributed and parallel data integration and mining and OLAP services, data provenance subsystem, etc.

• **CloudMiner** ([www.cloudminer.org](http://www.cloudminer.org))
  – includes the GridMiner‘s functionality.
  – additionally addresses data stream processing framework called StreamMiner.
ADMIRE Project

• The aim has been to create advanced, distributed data integration and analysis platform working on data streaming principles. Key outputs:
  – Language DISPEL (Data-Intensive Systems Process Engineering Language)
  – Process Designer (workflow construction/optimization)
  – ADMIRE Platform processing/executing DISPEL

• Suite of validating applications
  – customer relationship management
  – environmental risk management
  – analysing gene expression imaging data
  – seismology & astronomy
Development of Workflow Specifications – Concept of a Process Designer
DISPEL Example

use uk.org.ogsadai.SQLQuery;
use uk.org.ogsadai.TupleToWebRowSetCharArrays;
use eu.admire.Results;
use eu.admire.MultiClassify;

SQLQuery query = new SQLQuery;
String expression = "select * from stats limit 2000";
| expression -| => query.expression;
| "DbViennaTestResource" -| => query.resource;

TupleToWebRowSetCharArrays toWRS = new TupleToWebRowSetCharArrays;
query.data => toWRS.data;

Results result = new Results;
| "results" -| => result.name;
toWRS.result => result.input;

submit toWRS;

Remark: The ADMIRE Platform and DISPEL examples can be downloaded.
Process Designer‘s Modelling
GridMiner

• Full-fledged large-scale knowledge discovery system. It included
  – Mediator/wrapper data integration
  – Parallel and distributed OLAP, neural network and decision tree services
  – Interactive workflow management
  – Provenance subsystem, etc.

• Applications
  – Non-invasive blood glucose measurement
  – Management of TBI patients
Active Measurement

Data-file

Record

Measure

Send

Down-flow-key-point

Human-body-meridian

Up-flow-key-point

Up-flow point: lower electrical potential
Down-flow point: higher electrical potential
Fingers and toes: zero potential
CloudMiner

• GridMiner services
• Focus na stream management
  – elastic OLAP
  – stream mining
• New applications
  – Navigation systems
  – intelligent ambient assisted living
Bionic Global Navigation System Based on Polarized Light
Dementia Appl. Use Cases

Persons with dementia

Brain Stimulation

Orientation Support

Talking Key

Preventing dangerous situations

Finding Things again

Take me along!
Future Work: Next Generation of Ambient Assisted Living
Cloud-Based Ambient Asssisted Living
Data Space

• Primary data
  – Internal data: out- & indoor monitoring
  – External data: social networks, information associated with the map area, etc.

• Derived data
  – Knowledge patterns extracted
  – Rules (representing static knowledge)

• Background data
  – Different advices and observations
  – Pointers to knowledge sources, like literature, videos, etc.
Data Minig Tasks – Analytical Functions

• Observation/monitoring illness progression
  – Can data about the motion in an environment help discover/identify/predict patients’ worsening/impairment?
  – Estimating influence of the day time, weather, etc.

• Impact of data analysis results
  – Care modification/improvement Support for medical research (e.g. neuroscience)
  – Objective criteria for estimation of „Improvement of Quality of Life“.
Conclusions

• Growing importance of data-intensive research.
• Cloud features match the requirements
• CloudMiner/StreamMiner
• Future research plans: Intelligent Ambient Assisted Living
• Sky Computing (Metaclouds)?
Thank you for your attention!