DataMiningGrid project overview

Matevz Dolenc
Department of Construction Informatics
Faculty of Civil and Geodetic Engineering
University of Ljubljana
Project Details

- Falls under the Strategic Objective: “Grid-based Systems for Complex Problem Solving”
- Approximately 1.9 million EUR EU funding
- Start date 1st September 2004
- Website http://www.DataMiningGrid.org
Consortium Partners

University of Ulster (UU)

Fraunhofer Institute (FHG)

DaimlerChrysler (DC)

University of Ljubljana (LJU)

Technion Institute of Technology (TECH)
DataMiningGrid Drivers

- nutritional habits
- protein unfolding
- quality management
- gene reengineering
- text-mining
- modelling lake eco-systems
- grid monitoring
The Application Areas..

.. vary greatly in terms of their requirements:
• computationally expensive vs. large/distributed data sets
Motivation

**Thesis:**

- Knowledge industries/sectors generate fast-growing amounts of *data*
- *Automated* analysis and interpretation of these data (i.e. data mining) becoming a necessity

**Antithesis:**

- Processes, data, SW and resources increasingly *distributed* making data mining difficult

**Synthesis: DataMiningGrid**

- Facilitate *data mining* in distributed/grid computing environments
Vision

◆ Gap: No coherent framework for data mining in grid computing environments

◆ DataMiningGrid: First step towards a framework facilitating data mining in grid computing environments
  - Accessible data mining
  - Grid-aware data mining
  - Workflow-based mgt of data mining in grid computing environments
  - Industrial take-ups (demonstrators)
User Requirements

- Distributed and massive data
- Distributed operations
- Data types
- Data privacy and security
- Incorporating domain knowledge
- User-friendliness
- Resource identification and metadata
Evaluated Software

- GRIA, Condor, Unicore, Globus
- OGSA-DAI
- GridLab “upper-ware” Tools and Services
- Weka, Weka for WSRF, Grid-Weka implementations
- Orange, R, Lagramge,.. other legacy data mining algorithms and toolkits
- Triana Workflow Editor and Manager
- Fraunhofer Data Mining Editor and Manager,...
The **Open Grid Services Architecture (OGSA)**:
- defines a common, standard, and open architecture for grid-based applications
- it exploits stateful *Web Services*

**Web Services Resource Framework (WSRF, GGF)** is representation for current implementation of OGSA
- WSRF conforms to the WS-I (interoperability) profile
- standardizes basic behavior of grid services (the basic interfaces one commonly finds in a grid application)
Project’s Standardisation Work

✦ Areas of interest
  • MMML, PMML, CWM, XML for Analysis, GGF's DAIS and OGSA-DAI, WS-I+

✦ Standards bodies of interest
  • GGF, W3C, and COPRAS (CEN, CENELEC and ETSI)
Important Challenges/Risks

◆ Wide range of requirements arising from cross-sector data mining requirements
  • Text vs. relational data mining
  • Data vs. compute intensive mining
  • Interactive vs. batch operations
  • Privacy, access rights, firewalls, data types, platforms, languages, etc.

◆ Lack of suitable standards in both grid computing and data mining field

◆ Lack of stable/standardized grid middleware