Infrastructure, Data Cleansing and Mining for Scientific Simulations

Committee Members:

Dr. Bowyer
Dr. Flynn
Dr. Madey
Dr. Uhran
Agenda

- Overview
- Background
- Multi-tier infrastructure
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe
Overview

- Multi-tier infrastructure powers scientific simulations.
- Data cleansing algorithms result in better data quality.
- Data mining applications discover hidden knowledge in environmental and social science.
Motivation

Simulation Anytime and Anywhere

NOM

OSS

Infrastructure

Data Storage
Data Analysis
Reports

Collaboration
Personalization
Web-based
Agenda

- Overview
- Background
- Multi-tier infrastructure
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe
Background

- Projects under way
  - NOM
    - Research on natural organic matter (NOM)
    - Study evolution of NOM over time
    - Joint work of scientists across disciplines including chemists, biochemists, environmental scientists
  - OSS
    - Research on the open source software (OSS) development phenomenon
    - Study the behavior of OSS developers and their motivations
    - Joint work with social scientists
Simulation Models

- Standalone or traditional client-server
  - Software needs to be installed on clients
  - Incompatibility makes installation difficult

- Web-based using applets
  - Security – file permission, firewall
  - Inconvenience – plug-ins download
  - Network traffic – download before executing
  - Incompatibility – Swarm

- What should be done?
  - Web-based server-side simulation models
  - Centralized simulation management
  - Collaboration and personalization
Data Cleansing

- **Known approaches**
  - Sorted neighborhood (Stolfo 1995/1998)
    - Domain dependent keys for sorting
  - Record matching (Monge, 2000)
    - Edit distance only
  - String mapping (Li, 2003)
    - Potential high dimensional target space

- **Our approaches**
  - Sample database
  - Lipschitz mapping
Data Mining

- Data mining in astronomy
  - SKYCAT: star/galaxy classification (Fayyad, 1996)
  - JARTool: detect volcanoes on Venus (Burl, 1998)
  - Sapphire: find galaxies (Kamath, 2001)

- Data mining in biology
  - Bioinformatics
  - SARS diagnosis (ehealth.org)

- What should be done?
  - Data mining for social science (OSS)
  - Data mining for environmental science (NOM)
  - Add intelligence to simulation models by applying data mining results
Agenda

- Overview
- Background
- Multi-tier infrastructure
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe
The Simulation Manager

External Servers and Clients

Internet

Private network

Network Switch

Physical Layout
Multi-tier Architecture

HTTP Client tier

Client 1
Client 2
...
Client N

HTTP Server tier

Server

Application Server tier

Server 1
Server 2
Server 3
Server 4
Server 5

Database Server tier

SD
STDBY
SM
DW
Two Features

- **Load-balancing**
  - Scalability achieved
  - Implementation using JMS, AQ & EJB
  - Implementation using Shell scripts & PL/SQL

- **Simulation-resuming**
  - Reliability achieved
  - Checkpoint
  - Implementation using JTA/JTS
Load-balancing Using JMS & AQ

<table>
<thead>
<tr>
<th>job_id</th>
<th>resumed</th>
<th>checkpoint</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Job queue

Topic 1

mdb1 ➔ mdb2 ➔ mdb3 ➔ mdb4 ➔ mdb5 ➔ Loadavg queue ➔ Judge bean ➔ Topic 2

mdb6 ➔ mdb7 ➔ mdb8 ➔ mdb9 ➔ mdb10

Invoke simulation & update job queue
Shell Scripts & PL/SQL

- Dispatcher (HTTP server)
  - Dispatch simulations
  - Send KEEPALIVE messages to running simulations

- Intelligent agent (application server)
  - Upload load averages
  - Check simulations
  - Send ACK to KEEPALIVE messages
Load-balancing Algorithm

- Instance learning approach
  - Based on completion time prediction
- Two step completion time prediction
  - Completion time estimation
    - Load average
    - Data amount
  - Completion time prediction
    - Nearest neighborhood
Completion Time Estimation

 Completion time estimation formula
Checkpoint

JDBC

JTA/JTS

One transaction
Checkpoint Issues

- Checkpoint data
  - All data for restarting the simulation
  - Size depends on number of agents

- Checkpoint frequency
  - Checkpoint-interval
    - # of MB data
  - Checkpoint-timeout
    - # of minutes
Simulation-resuming

◆ To restart a terminated simulation
  - A new simulation with same job_id inserted into the job queue
  - A terminated simulation has smaller job_id than new simulations, higher priority

◆ In case of application server failure
  - All simulations’ job_ids inserted into the job queue
  - All simulations will be running on other application servers
Collaboration Suite

NOM

NOM Discussion Board

NOM BBS
NOM researchers can discuss modeling, programming, reporting and so on.

NOM Simulator
NOM researchers can start simulations and view information such as reports for their simulations.

NOM Chat Room

Your Name: 
Password: 
Member [ ] Not Member [ ]
Register
Visitor: 
go!
Graphical Reports

Please provide the following information.
Current Time Step is: 4799

Time Step: 4799
Status: adsorbed
Start Y: 0
End Y: 299

Generate Reports
Generate XML file

NOM Adsorption Simulation

Weight Distribution

<table>
<thead>
<tr>
<th>weight</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>380</td>
<td>0.05</td>
</tr>
<tr>
<td>732</td>
<td>0.25</td>
</tr>
<tr>
<td>1464</td>
<td>0.30</td>
</tr>
<tr>
<td>2928</td>
<td>0.15</td>
</tr>
<tr>
<td>6866</td>
<td>0.10</td>
</tr>
</tbody>
</table>
XML Reports

Please provide the following information.
Current Time Step is: 4799 Refresh

<table>
<thead>
<tr>
<th>Time Step</th>
<th>Status</th>
<th>Start Y</th>
<th>End Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>4799</td>
<td>adsorbed</td>
<td>0</td>
<td>299</td>
</tr>
</tbody>
</table>

Weight Distribution

<table>
<thead>
<tr>
<th>weight</th>
<th>count</th>
<th>total</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>380</td>
<td>8</td>
<td>76</td>
<td>0.1053</td>
</tr>
<tr>
<td>732</td>
<td>21</td>
<td>76</td>
<td>0.2763</td>
</tr>
<tr>
<td>1464</td>
<td>25</td>
<td>76</td>
<td>0.3289</td>
</tr>
<tr>
<td>2928</td>
<td>13</td>
<td>76</td>
<td>0.1711</td>
</tr>
<tr>
<td>5856</td>
<td>9</td>
<td>76</td>
<td>0.1184</td>
</tr>
</tbody>
</table>

Generate Reports
Generate XML file
Agenda

- Overview
- Background
- Multi-tier information system
- **Data mining applications**
- Summarize
- Timeframe
Methodology

- Traditional approach
  - Form hypotheses
  - Verify hypotheses by finding patterns in data

- Data mining approach
  - Find patterns in data
  - Form hypotheses
  - Design simulation models
  - Verify hypotheses

- U. Fayyad, J. Gray at Microsoft Research
Technology & Software

- Data mining technology
  - Clustering
    - K-means
    - Orthogonal cluster
  - Classification
    - Decision tree
    - Naïve Bayes
  - Association rules
    - Apriori

- Data mining software
  - Oracle Data Mining Suite
  - DM4J
  - JDeveloper
Study behavior of open source software (OSS) developers
- Agent-based
- Stochastic

Data mining involving
- Clustering
- Classification
  - Churn prediction
  - Acquisition prediction
- Association rules
OSS Data Warehousing

- Data from sourceforge.com
  - Developers
  - Projects

- Data warehousing
  - Table partitioning
  - Aggregation
  - Star schema
  - Analysis SQL
  - ETL tools → Warehouse Builder
Study behavior of natural organic matter (NOM)
- Agent-based
- Stochastic

Data mining involving
- Clustering
  - Micelle formation
- Classification
  - Transportation prediction
  - Adsorption prediction
- Association rules
Agenda

- Overview
- Background
- Multi-tier information system
- Data mining applications
- Summarize
- Timeframe
Summarize

- Multi-tier information system integrates
  - Application servers & reports server
  - Database servers
  - Data warehousing & data mining
  - Swarm
- Collaboration suite
- Data mining guided model-design
Insights & Impacts

- Server-side simulation models
  - Centralized simulation management
  - Centralized data repository

- Collaboration suite
  - Simulation sharing
  - Knowledge sharing

- Data mining applications
  - Find patterns in data
  - Model deployment for simulation-design
Agenda

- Overview
- Background
- Multi-tier information system
- Data mining applications
- Summarize
- Timeframe
Timeframe

May 2003 ~ May 2004

Implement infrastructure

Data collection & statistical analysis

Data mining model design

Data mining model evaluation

Deployment

Writing up
Expected Publications

- Information system design for scientific simulations
  - By August 2003
- Data warehousing for scientific simulations
  - By November 2003
- Data mining for OSS
  - By February 2004
- Data mining for NOM
  - By March 2004
Demo

Demonstration
Finally

Thank you!
Features

- Multi-tier information system
  - HTTP client tier → HTTP server tier → Application server tier → EIS tier
- Scalability at the application server tier
  - Load-balancing
- Reliability at the application server tier
  - Simulation-resuming
- Reliability at the database tier
  - Standby databases
Features (cont.)

- Data mining models
  - Stored in database
  - Stored Java procedures
  - PL/SQL procedure call using JDBC

- Simulation models
  - Agent-based
  - Stochastic
  - Data mining guided