Infrastructure, Query Optimization, Data Warehousing and Data Mining in Support of Scientific Simulation

Yingping Huang

Department of Computer Science and Engineering
University of Notre Dame

Tuesday, October 29, 2002

Partially supported by NFS-ITR
Route

- Research area, results & motivation
- Background & technologies
- Modeling & simulation
- Infrastructure
- GUI & web interface
- Query optimization
- Data warehousing
- Data mining
- Summary & future work
Research Area and Results

- The domain
  - Scientific simulation
    - Natural organic matter (NOM)
    - Environmental biocomplexity
- The results: A simulation model
  - Agent-based
  - Stochastic
  - Web-based: J2EE & Oracle
  - Load-balancing and fail-over enabled
  - Data warehousing & data mining features included
Motivation

- IT: A fourth paradigm of scientific study? (J. Gray, et al, 2002; Fox, 2002)
  - Three previous approaches to scientific research:
    - Observation & theory
    - Hypothesis & experiment
    - Computational X & simulation
  - Information technologies
    - J2EE & middleware & XML
    - Databases & Data Warehouses
    - Data Mining
    - Visualization
    - Statistical analysis
- Natural organic matter (NOM)
Technology Used

- Agent-based modeling
  - SWARM: a library
- Stochastic modeling
- J2EE
  - JSP
  - Servlet
  - EJB
- Application Server
- Oracle
  - RDBMS
  - JDBC
  - PL/SQL
  - Reports Server
  - Data Warehouse
  - Data Mining
Agent-based Modeling

- Property of intelligent agents
  - Autonomous behavior
  - Individual world of view
  - Communicative & cooperative capacity
  - Intelligent behavior
  - Spatial mobility
- De-central control
  - Social insects & birds
- Emergent behavior
  - Patterns, clusters, self organization, etc
Chemical Reactions Models

Classification criteria

- Simulation time: discrete or continuous
  - Computers only do discrete computations
- State-space: discrete or continuous
  - n-dimensional space containing all states of n variables
- Evolution of system: deterministic or stochastic
  - Deterministic: State of system completely specified at all times
  - Stochastic: State of system represented by probability distributions & Evolution determined by probability events
Simulation of NOM and Microbial-Environmental Interactions

- NSF - ITR - Division of Environmental Biology
- Interdisciplinary project
  - Chemist
  - Geomicrobiologist
  - Biologist
  - Ecologist
  - Computer Scientist
- Stochastic Simulation of Environmental Transformations of Natural Organic Matter
  - In soil
  - In solution
NOM is ubiquitous in terrestrial, aquatic and marine ecosystems
  - Results from breakdown of animal & plant material in the environment

Important role in processes such as
  - compositional evolution and fertility of soil
  - mobility and transport of pollutants
  - availability of nutrients for microorganisms and plant communities
  - growth and dissolution of minerals

Important to drinking water systems
  - Impacts drinking water treatment
  - Impacts quality of well water
Natural Organic Matter (cont)

Hardwood Swamp
Natural Organic Matter (cont)

Open Channel
Natural Organic Matter (cont)
Compositional evolution of NOM is an interesting problem. Important aspect of predictive environmental modeling. Prior modeling work is often too simplistic to represent the heterogeneous structure of NOM and its complex behaviors in ecosystems (e.g., carbon cycling models) or too compute-intensive to be useful for large-scale environmental simulations (e.g., molecular models employing connectivity maps or electron densities).

Hence, a Middle Computational Approach is taken …
  – Agent-based & stochastic
Previous work

- Models developed by other researchers
  - Deterministic models
    - METASIM (Park & Wright, 1973)
    - SCAMP (Saura, 1993)
  - Stochastic models
    - CKS (IBM, 1995)
    - BESS (Punch, 1997)
    - STOCHSIM (Firth & Bray, 2001)
Our Model

- Agent-based stochastic simulation
- GUI Version - Stand Alone
  - Animation of molecules
- Web-Based Version
  - OC4J/Orion Server & Oracle Reports
  - Oracle database servers
- Load-balancing & fail-over
  - Goal: efficiency, availability & reliability
- Data warehousing & Data Mining
  - Goal: data/pattern analysis
Object oriented: Molecules and microbes are objects
  - Molecules and microbes have attributes
    - Heterogeneous mixture: different attributes
  - Molecules have behaviors (physical & chemical processes)
    - Behaviors are stochastically determined
    - Dependent on the:
      - Attributes (intrinsic parameters)
      - Environment (extrinsic parameters)
Objects of interest
- Macromolecular precursors: large molecules
  - Cellulose
  - Proteins
  - Lignin
- Micromolecules: smaller molecules
  - Sugars
  - Amino acids
- Microbes
  - Bacteria
  - Fungi
Attributes

- Elemental composition
  - Number of C, H, O, N, S and P atoms in molecule
- Functional group counts
  - Double-bonds
  - Ring structures
  - Phenyl groups
  - Alcohols
  - Phenols, ethers, esters, ketones, aldehydes, acids, aryl acids, amines, amides, thioethers, thiols, phosphoesters, phosphates
- The time the molecule entered the system
- Precursor type of molecule
  - Cellulose, protein, lignin, etc
Modeling (cont)

Behaviors (reactions and processes)
  - Physical processes
    - Adsorption (stick) to mineral surfaces
    - Aggregation/micelle formation
    - Transport downstream (surface water)
    - Transport through porous media
  - Chemical reactions
    - Abiotic bulk reactions: free molecules
    - Abiotic surface reactions: adsorbed molecules
    - Extracellular enzyme reactions on large molecules
    - Microbial uptake by small molecules
Modeling (cont)

- Environmental parameters
  - Temperature
  - pH
  - Light intensity
  - Simulation time
  - Microbial activity
  - Water flow rate/pressure gradient
  - Oxygen density
A Molecule at a Time Step

1. Read probability table
2. A random number is generated
3. First order reaction?
   - Yes: Do first order reaction
   - No: Find the second molecule
4. Do second order reaction
5. Update the probability table
UML Use Case Diagram

ObserverSwarm

- Update Probe Display
  - <<executes>>

- Update Schedule
  - <<executes>>

- Execute GUI
  - <<uses>>

- Execute Simulation Schedule
  - <<uses>>

ModelSwarm

- Move to New location
  - <<extends>>

- Update ProbabilityTable
  - <<uses>>

- Update Molecule
  - <<uses>>

- Update World
  - <<uses>>

- Write Database
  - <<uses>>

- Update World Display
  - <<uses>>
GUI Animation

Black - No Adsorption
Gray - Levels of Adsorption
Red - Lignins
Blue - Proteins
Green - Cellulose
Yellow - Reacted
Orange - Adsorbed
NOM Simulator

Welcome to NOM Research Group!

You must sign in to use the simulator!

Existing Users
Enter your userid and password to sign in

Userid: 
Password: 

Sign In

New users? Sign up now
NOM Simulator

Welcome to NOM Research Group 8 Y

NOM Simulator: Reports

Currently, you have the following sessions invoked. The first one is your most recent session. You can view reports for each session by click the following links. To start a new simulation, click here. To cleanup terminated sessions, click here.

- Session 118: Terminate Session
  - Reactions Reports
- Session 117: TERMINATED
  - Reactions Reports
- Session 116: TERMINATED
  - Reactions Reports
- Session 115: Terminate Session
  - Reactions Reports
- Session 114: Terminate Session
  - Reactions Reports
- Session 113: TERMINATED
  - Reactions Reports
- Session 112: Terminate Session
  - Reactions Reports
- Session 111: Terminate Session
  - Reactions Reports
- Session 110: Terminate Session
  - Reactions Reports
**NOM Simulator**

Welcome to NOM Research Group

<table>
<thead>
<tr>
<th>Introduction</th>
<th>Environment</th>
<th>Molecules</th>
<th>Summary</th>
</tr>
</thead>
</table>

**NOM Simulator: Introduction**

To properly use the simulator, we need to gather data for environment and molecule types.

The wizard will walk you through several tasks:

- Provide environment variables. If you provided environment variables before, we will retrieve your information to let you edit.
- Provide molecule types and number of molecules of this type. You can also edit and delete your saved molecule information.
- Invoke the simulation
### NOM Simulator: Environment

<table>
<thead>
<tr>
<th>Simulation Time(days)</th>
<th>Microbe Density</th>
<th>pH Value</th>
<th>PKW</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.00010</td>
<td>7.0</td>
<td>54.0</td>
</tr>
<tr>
<td>Fungal Density</td>
<td>0.0010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxygen</td>
<td>3.0e-4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Environment Information**
Please provide the environment variables for your simulation. You may also edit your environment variables here. Before submitting the form, please make sure that all the fields must be integers or doubles. If you have already provided environment variables, you may choose to skip this step.

[Skip & Next]
[Save & Next]

Step 2 of 4
# NOM Simulator

Welcome to NOM Research Group & Y

## NOM Simulator: Molecule

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Cellulose</th>
<th>Lignin</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Atom) C:</td>
<td>360</td>
<td>400</td>
<td>240</td>
</tr>
<tr>
<td>(Atom) H:</td>
<td>602</td>
<td>322</td>
<td>332</td>
</tr>
<tr>
<td>(Atom) N:</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>(Atom) O:</td>
<td>301</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>(Atom) S:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(Atom) P:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Double Bond:</td>
<td>60</td>
<td>199</td>
<td>59</td>
</tr>
<tr>
<td>Total Ring Structures</td>
<td>60</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>Phenyl Groups:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alcohol:</td>
<td>182</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Phenols:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethers:</td>
<td>119</td>
<td>118</td>
<td>0</td>
</tr>
<tr>
<td>Esters:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ketones:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aldehydes:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acids:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amines:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Known Molecule Information**

There are three types of already defined Molecule, please give the percentage of each. Give a value of 0 for percentage if you don't want to include this molecule type in your simulation. If you do not want to include any of these three types of molecules, you may click the Skip & Next button, otherwise, please click the Save & Next button. Default values are 0.
<table>
<thead>
<tr>
<th>Molecular Groups</th>
<th>Percentage 0</th>
<th>Percentage 1</th>
<th>Percentage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohols</td>
<td>0.182</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Phenols</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethers</td>
<td>0.119</td>
<td>118</td>
<td>0</td>
</tr>
<tr>
<td>Esters</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ketones</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aldehydes</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Acids</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Aryl Acid</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amines</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ring N.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amines</td>
<td>0</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>Thioethers</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Thiols</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phosphoesters</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H-phosphoesters</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phosphates</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Percentage:
- 0.0%
- 0.0%
- 0.0%

The Save & Next button. Default values are 0.

Your saved molecules:

<table>
<thead>
<tr>
<th>Molecule Name</th>
<th>Percentage</th>
<th>Edit or Delete?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>0.05</td>
<td>Delete</td>
</tr>
<tr>
<td>Cellulose</td>
<td>0.33</td>
<td>Delete</td>
</tr>
<tr>
<td>Lignin</td>
<td>0.33</td>
<td>Delete</td>
</tr>
</tbody>
</table>
# NOM Simulator

Welcome to NOM Research Group 2 Y

## NOM Simulator: Molecule

<table>
<thead>
<tr>
<th>Molecule Name</th>
<th>Percentage</th>
<th>Molecule Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>29.0</td>
<td>Please provide molecule's name, percentage, number of atoms of molecules for your simulation. Please remember, except &quot;Molecule Name&quot;, all fields should be integers or doubles. &quot;Percentage&quot; should be between 0 and 100.</td>
</tr>
<tr>
<td>(Atom) C</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(Atom) N</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(Atom) S</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(Atom) H</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(Atom) O</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(Atom) P</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Doublebond</th>
<th>0</th>
<th>Functional Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rings</td>
<td>0</td>
<td>Please provide a number for each functional group. Default value is 0.</td>
</tr>
<tr>
<td>Phenyl</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Alcohol</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Arylacids</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Amines</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>RingN</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Amides</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
## NOM Simulator: Summary

We have gathered all information we need, you may invoke your simulation now. **Invoke Simulation**

### Environment Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation Time</td>
<td>2.0</td>
</tr>
<tr>
<td>Microbe Density</td>
<td>0.0010</td>
</tr>
<tr>
<td>Fungal Density</td>
<td>0.0010</td>
</tr>
<tr>
<td>pH Value</td>
<td>7.0</td>
</tr>
<tr>
<td>Temperature</td>
<td>300.0</td>
</tr>
<tr>
<td>PKW</td>
<td>14.0</td>
</tr>
<tr>
<td>Oxygen Density</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Light Density</td>
<td>4.0E-6</td>
</tr>
</tbody>
</table>

### Molecule Information

<table>
<thead>
<tr>
<th>Molecule Name</th>
<th>Percentage Edit or Delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>5.0 Delete</td>
</tr>
<tr>
<td>Cellulose</td>
<td>33.0 Delete</td>
</tr>
<tr>
<td>Lignin</td>
<td>33.0 Delete</td>
</tr>
<tr>
<td>Molecule A</td>
<td>29.0 Delete</td>
</tr>
</tbody>
</table>
Welcome to NOM Research Group X Y

NOM Simulator: Reports

Currently, you have the following sessions invoked. The first one is your most recent session. You can view reports for each session by clicking the following links. To start a new simulation, click here. To cleanup terminated sessions, click here.

- Session 119: Terminate Session
  - Reactions Reports
- Session 118: Terminate Session
  - Reactions Reports
- Session 117: Terminated
  - Reactions Reports
- Session 116: Terminated
  - Reactions Reports
- Session 115: Terminate Session
  - Reactions Reports
- Session 114: Terminate Session
  - Reactions Reports
- Session 113: Terminated
  - Reactions Reports
- Session 112: Terminate Session
  - Reactions Reports
- Session 111: Terminate Session
  - Reactions Reports
Reactions By Type

Reactions vs Time

Welcome to NOM Research Group's NOM Simulator: Reports
The Simulation Infrastructure

Remote Clients/Servers

Application Servers/Simulation Running

Internet

gemini.cse.nd.edu
Intel Dual 400
Win2K Server
OC4J/Orion Server
Reports Server

Web Interface/Reports

HTTP

Joy.cse.nd.edu
Intel Dual 800
Redhat 7.2
OC4J/Orion Server

Tenor.cse.nd.edu
Intel Dual 800
Redhat 7.2
OC4J/Orion Server

Simu2.world
foy.t.cse.nd.edu
Intel Dual 400
Redhat 7.2
Oracle9i 2
Data Mining

Etch.world
Symphony.cselab
Intel Dual 733
Win2K Server
Oracle9i 2
Data Mining

Database Servers/Data Mining Servers

Mynom.world
bigband.cselab
Intel Dual 733
Solaris 8
Oracle8i

JDBC

HTTP

Internet

Redhat 7.2
OC4J/Orion Server
Loosely coupled distributed systems
- 2 Application servers (Orion Servers)
- 3 Database servers (Oracle)
- Reports server (OC4J Server/Reports Server)

Load balancing (round robin based on computational needs)
- application servers & database servers

Fail over
- application servers & database servers
- Multi-master replication of important tables

Why fail-over (Assume down probability $p$ for each machine)
- No fail-over
  - Simulation system down probability: $1-(1-p)^2 = 2p-p^2$
- With fail-over
  - Simulation system down probability: $1-(1-p^2)(1-p^3) = p^2 + p^3 - p^5$
- Improvement:
  - $2/p = 200$ if $p=0.01$ (the smaller $p$, the larger improvement)
Simulation Configuration Data Model

<<Persistent>>
Users

- ID: INTEGER
- FIRST_NAME: VARCHAR2
- LAST_NAME: VARCHAR2
- USERID: VARCHAR2
- PASSWORD: VARCHAR2
- EMAIL: VARCHAR2
- PHONE: VARCHAR2

<<Persistent>>
Environ

- SIMULATIONTIME: NUMBER
- MD: NUMBER
- FD: NUMBER
- PH: NUMBER
- T: NUMBER
- PKW: NUMBER
- O2: NUMBER
- I: NUMBER
- USER_ID: INTEGER

<<Persistent>>
MoleculeAttributes

- TYPEID: INTEGER
- C: INTEGER
- H: INTEGER
- N: INTEGER
- O: INTEGER
- S: INTEGER
- P: INTEGER
- DOUBLEBOND: INTEGER
- RINGS: INTEGER
- PHENYL: INTEGER
- ALCOHOLS: INTEGER
- PHENOLS: INTEGER
- ETHERS: INTEGER
- ESTERS: INTEGER
- KETONES: INTEGER
- ALDEHYDES: INTEGER
- ACIDS: INTEGER
- ARYLACIDS: INTEGER
- AMINES: INTEGER
- RINGN: INTEGER
- AMIDES: INTEGER
- THIOETHERS: INTEGER
- THIOLS: INTEGER
- PHOSPHOESTERS: INTEGER
- HPHOSPHOESTERS: INTEGER
- PHOSPHATES: INTEGER
- USER_ID: INTEGER
- NAME: VARCHAR2
- QUANTITY: NUMBER
Simulation Data

- **Molecule_ID**
  - All molecule entered the system or produced by chemical reactions have a molecule_id

- **Session_ID**
  - Each simulation session has a unique ID

- **TimeStamp**
  - Each time step of the system is associated with molecules

- **xPos & yPos**
Simulation Data (Cont)

- Parent1 & Parent2
  - If first order reaction, parent2 is NULL
- Reaction probabilities
  - After a chemical reaction, probability tables are updated
- Molecule structures
  - After a chemical reaction, molecule structures are updated
Query Optimization

- Insertion performance
  - Disable indexes
  - Disable constraints
- Query performance
  - Indexes
  - Aggregation tables
- Space utilization
  - PCTFREE & PCTUSED & INITRANS & MAXTRANS
  - Drop indexes
Query/Report Examples

Example 1:
- Show the number of chemical reactions for each of the ten reaction types so far in the simulation using bar charts

Example 2:
- Create a line graph which shows the trend of the total number of chemical reactions vs time steps.
Example 1

SQL> select nom.reactiontype "Reaction Type",
2 reactiontype.rname "Reaction Name",
3 count(nom.moleculeid) "Reactions"
4 from nom, reactiontype
5 where nom.reactiontype=reactiontype.rtype
6 and sessionid=:session_id and user_id=:user_id
7 group by nom.reactiontype, reactiontype.rname
8 order by nom.reactiontype;

Elapsed: 00:00:10.03
Example 2

SQL> select t1.timestamp “Time Step”,
2       sum(t2.total) “Reactions”
3   from (select timestamp,
4           count(moleculeid) total
5          from nom
6         where sessionid=:session_id
7             and user_id=:user_id
8          group by timestamp ) t1,
9       (select timestamp,
10          count(moleculeid) total
11          from nom
12        where sessionid=:session_id
13            and user_id=:user_id
14          group by timestamp ) t2
15  where t2.timestamp <= t1.timestamp
16 group by t1.timestamp;

Elapsed: 01:20:10.23
Aggregation Tables

Example 1
- **REACTIONS_BY_TYPE**
  - Session_ID & Reaction Type & Reactions
  - Updated at the end of every time step

Example 2
- **REACTIONS_BY_TIME**
  - Session_ID & Time Step & Total Reactions
  - A new row inserted at the end of every time step
## Insertion and Query Performance Comparison

<table>
<thead>
<tr>
<th>Scenario (&gt;16million)</th>
<th>Insertion (sec/row)</th>
<th>Query Time (example 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No indexes No aggregation</td>
<td>0.0106</td>
<td>&gt;1 hour</td>
</tr>
<tr>
<td>With indexes</td>
<td>0.0122</td>
<td>&gt;0.5 hour</td>
</tr>
<tr>
<td>With aggregations</td>
<td>0.0107</td>
<td>5 seconds</td>
</tr>
</tbody>
</table>
Data Warehousing

A data warehouse is a database with the following properties:

- **Subject oriented**
  - Define data warehouse by subject matter
- **Integrated**
  - Consistent format, data integrity
- **Non-volatile**
  - Rarely update
- **Time-variant**
  - Data collected over time, temporal attributes

Inmon, 1996
Logical Design of The Data Warehouse

- Conceptual & abstract
  - Define the metadata
  - Entity-relationship modeling
  - Using Oracle Designer to generate DDL

- Two design approaches
  - Detail and Summary Schema
  - Star Schema
Detail and Summary Schema

- Summary Chemical Reactions By pH and User
- Summary Chemical Reactions By pH and Session
- Summary Chemical Reactions By pH
- Summary Chemical Reactions By Temperature
- Summary Chemical Reactions By Reaction Type
- Summary Chemical Reactions By Reaction Type And Time Stamp

Detailed Simulation Data For Each Session
Advantages and Disadvantages of Detail and Summary Schema

Advantages
- Easy to navigate
  - Incorporate data from other related tables to avoid join operation from the summary
  - For example, The REACTIONS_BY_TYPE avoids join of NOM and REACTIONTYPE.

Disadvantages
- What summarizations are anticipated?
Star Schema

- Derived from multidimensional database design (Kimball, 1996)
- Facts tables
  - Central large tables
- Dimension tables
  - Descriptive attributes about a dimension in facts tables
- Fact table has a foreign key relationship to each dimension table
- More flexible than Detail and Summary Schema
  - Summary and GROUP BY in Detail and Summary Schema
A Star Schema

**USERS DIMENSION**
- user_id
- first_name
- last_name
- phone
- email
- password

**SESSIONS DIMENSION**
- session_id
- user_id
- sid
- status
- expected

**MOLECULES DIMENSION**
- molecule_id
- c
- h
- doublebond
- amines
- prob_0
  - ...

**ENVIRONMENT DIMENSION**
- environ_id
- temperature
- md
- fd
- pH
  - ...

**REACTIONS**
- user_id
- session_id
- molecule_id
- reaction_type
- environment_id
- xpos
- ypos
- timestamp

**REACTIONTYPE DIMENSION**
- reaction_type
- reaction_name
Build the Data Warehouse

- Oracle database as the data warehouse
- Tablespaces design
  - I/O contention reduction
    - Files associated with each tablespace are striped across multiple disks
- Predict space requirement
  - Load sample data
  - ANALYZE command
  - STATSPACK
- Space availability insurance
  - AUTOEXTEND
- Partitioned tables and indexes
Populate the Data Warehouse

- **Tools**
  - SQL*Loader
  - Export/Import
  - SQL*Plus copy command
  - CREATE TABLE ... AS SELECT command
  - JDBC

- **Data preprocessing**
  - Data Cleansing
  - Resolve name and format inconsistencies

- **Summary & aggregation**
Query Optimization for the Data Warehouse

- Optimization techniques involved
  - Ordered hint: SELECT /*+ordered*/ ...
    - reducing parsing time
    - For example, join of 9 tables has 8!=40320 join combinations; parsing takes more than 30 minutes
  - Star hint: SELECT /*+star*/ ...
    - Hash join
    - Bitmap indexes
    - Result in reducing I/O
  - Partitioning
    - Join devided into small joins
Data mining refers to extracting or mining knowledge from a large amount of data.

Other terms:
- Knowledge discovery in database (KDD)
- Data/pattern analysis
- Information retrieval
- Machine learning
Data Mining as Step of KDD

Data Mining

Data Warehouse

Knowledge

Evaluation and Presentation

Patterns

Extracting, Transforming and Loading

Databases
Oracle Data Mining

- ODM has two components
  - Data Mining API
    - Provides an early look at concepts and approaches begin proposed for the emerging standard Java Data Mining (JDM)
    - Based on data mining standards
      - Object Management Group’s Common Warehouse Metadata (CWM)
      - Data Mining Group’s Predictive Model Markup Language (PMML)
      - International Standards Organization’s SQL/MM for Data Mining
  - Data Mining Server
    - Server-side in-database component that performs data mining
Data Mining Functions

- Classification (supervised)
  - Naïve Bayes algorithm
  - Decision tree algorithm: CART & C5.0
- Clustering (unsupervised)
  - Low inter-cluster similarity
  - High intra-cluster similarity
- Association Rules (unsupervised)
- Attribute Importance (supervised)
Data Mining Steps

- **Build model**
  - Build model using training set
- **Test model**
  - Data has same format as model-build data
- **Compute lift (if applicable)**
  - Usually for classification
  - To test whether the model is useful
- **Apply model**
  - Data has same format as model-build data
Model-Build Process

Data Warehouse

- Define mining Function Settings
- Mining function Settings
- Data Preprocessing
- Core Algorithm
- Internal Bin Boundary Tables
- Model

- Bin Boundary Tables
- Model Build Data Table
Model-Apply Process

Data Warehouse

Bin Boundary Tables

Model Apply Data Table

Apply Output

Specify Apply Output

Data Preprocessing

Core Algorithm (Apply)

Model Internal Bin Boundary Tables

Result Table
Clustering Algorithms

- **Partitioning**
  - K-means: each cluster represented by the mean value of the objects in the cluster
  - K-medoids: each cluster represented by one of the objects located near the center of the cluster

- **Hierarchical**
  - Agglomerative: bottom-up
  - Divisive: top-down
Clustering Algorithms (cont)

- **Density-based**
  - Continuing growing cluster as long as the density in the neighborhood exceeds a threshold

- **Grid-based**
  - Quantize the object space into finitely many cells that form a grid structure
  - Fast processing time

- **Model-based**
  - Statistic approach
  - Neural network approach
Oracle Clustering Algorithms

- **Enhanced k-means algorithm**
  - Hierarchical k-means algorithm
  - Top-down approach
  - The cluster with largest distortion (sum of distances to the cluster center) is split until desired number of clusters reached

- **O-Cluster algorithm**
  - Grid-based
  - Hierarchical
  - A unit (cell) is dense if the density exceeds SENSITIVITY
Build Clustering model for Data Warehouse

- Clustering model build steps
  - Data is standardized
  - Connect to the data mining server
  - Create a PhysicalDataSpecification object for model build data
  - Create a MiningFunctionSettings object which specifies the algorithm settings
  - Build the model
Apply Clustering Model to Data Warehouse

Programming steps
- Model apply data is standardized
- Connect to the data mining server
- Create a PhysicalDataSpecification object for input data which is the data to be scored
- Create a LocationAccessData object for output data, which is the table to store the scoring results
- Create a MiningApplyOutput object for output data, capturing the format of output
- Score the data
Apply Clustering

- Model-build data format
  - A table POINTS with attributes x & y
    - Points are chosen from the data warehouse
    - Standardized: x & y are in [0,1)
    - 16 million records

- Clusters explanation
  - Dense areas in soil or solution
  - Emerging behavior of random molecules
Aggregation & Micelle Formation

Maurice, 1999

NOM Rings
## Comparison of Enhanced k-means and O-Cluster

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Build Time (16M rows)</th>
<th>Cluster Shape</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced k-means</td>
<td>34 min</td>
<td>Spherical</td>
<td>8 (specified)</td>
</tr>
<tr>
<td>O-Cluster</td>
<td>14 min</td>
<td>Rectangular</td>
<td>15 (auto)</td>
</tr>
</tbody>
</table>
Summary

Contributions are

- New model which treats NOM as a heterogeneous mixture
- Simulation system with advanced web & database tools
- System aspects of implementation of load-balancing and fail-over
- Basic data mining features
Future Work

- Simulation system
  - More features
  - Reliability
  - Efficiency
  - Intelligent simulation configuration wizards

- Simulation data analysis
  - More data mining algorithms
  - Ad hoc queries

- Collaboration tools
  - Oracle Collaboration Suite