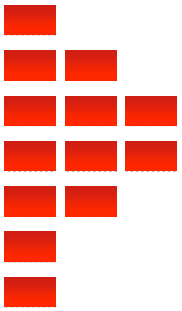
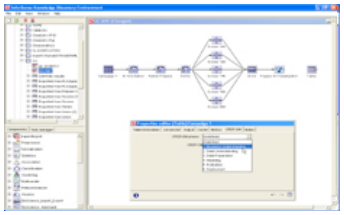
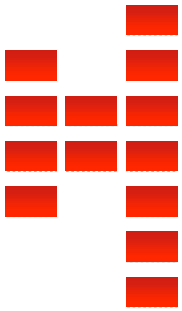
Rapid Application Development using InforSense Open Workflow and Daylight Technologies



InforSense

Deliver Discovery Value





Anthony Arvanites

Daylight User Group Meeting

March 10, 2005

Outline

InforSense

1. Company Introduction
2. What Does InforSense Do?
3. Daylight DayCart Integration
4. Demos
   - Chemical Warehouse
   - Modeling Application

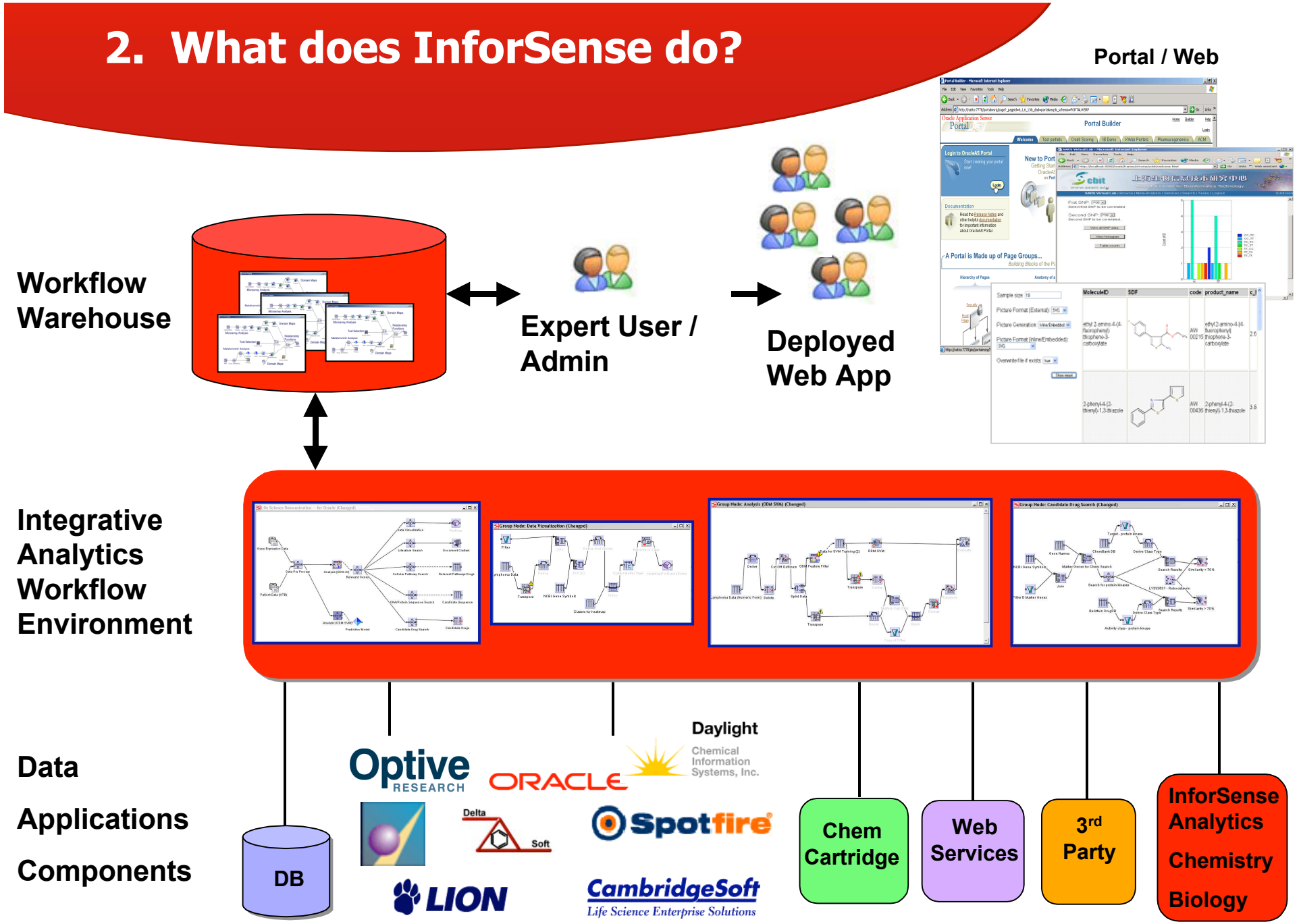
1. Company Introduction

InforSense

- Company founded Nov 1999 to commercialize Imperial College Super Computing & Data Mining IP
- 45 staff plus 20 developers in Shanghai
- 4 out of top 15 pharma are customers

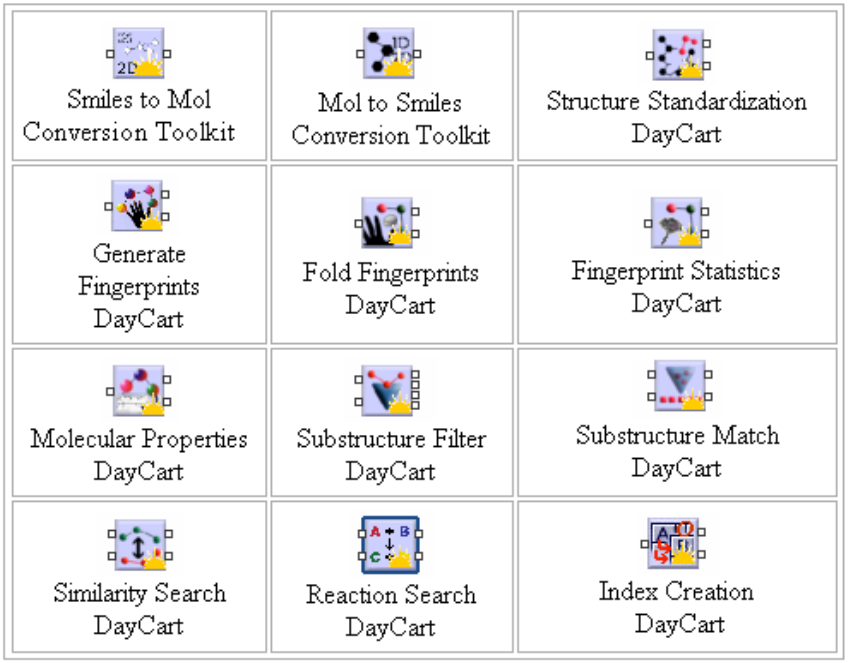
- Workflow centric integrative analytics infrastructure
- Integrate cross-domain data, applications, components into process workflow without programming
- Capture, manage, deploy & audit analytic workflow processes

2. What does InforSense do?



DayLight Functions Integrated into KDE

InforSense

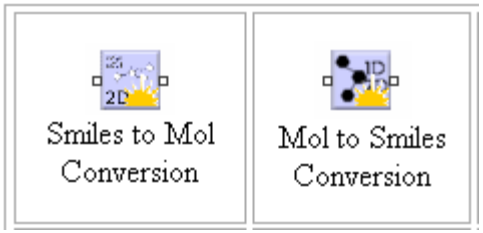


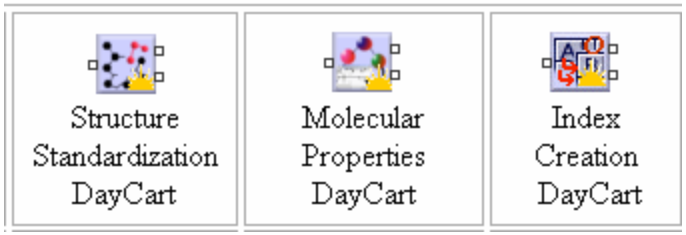
- **Daylight 'Contrib' Code**
  - Similes2Mol and Mol2Smiles
- **DayCart Components**
  - Exact Match, Substructure, Similarity, Structure Standardization, Molecular Properties, Fingerprint, Fingerprint Folding and Statistics, Reaction Handling, and Index Creation

- Installed & created prototype within 8 hours with existing oracle components
- 1 week development time based upon specs

Standardization & Indexing

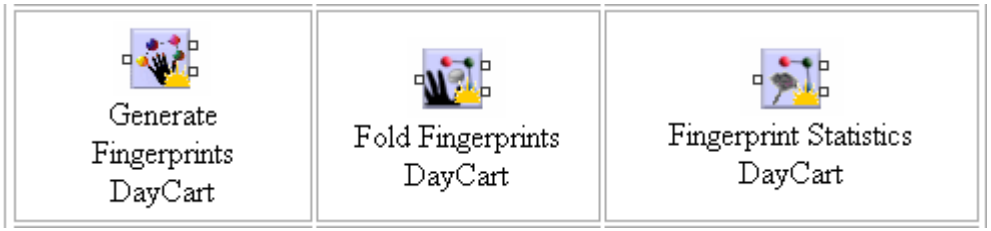
InforSense





- **Daylight 'Contrib code' conversion utilities**
  - MOL2SMI and SMI2MOL performs conversion to and from SD and MOL formats.
- **Structure Standardization**
  - Operators smi2cansm, vcs_desalt and vcs_normalize
  - Remove molecular fragments found in c$dcischem.salts
  - SMIRKS-based structure normalization on input Smiles based on c$dcischem.transform
- **Molecular Properties**
  - Operators smi2netch, smi2hcount, smi2mf, smi2amw
  - Access Dayprop via Dayproptalk from DayCart
- **Index Creation**
  - Creates indexes ddexact, ddblob, ddgraph, or ddrole

DayCart Fingerprint Components



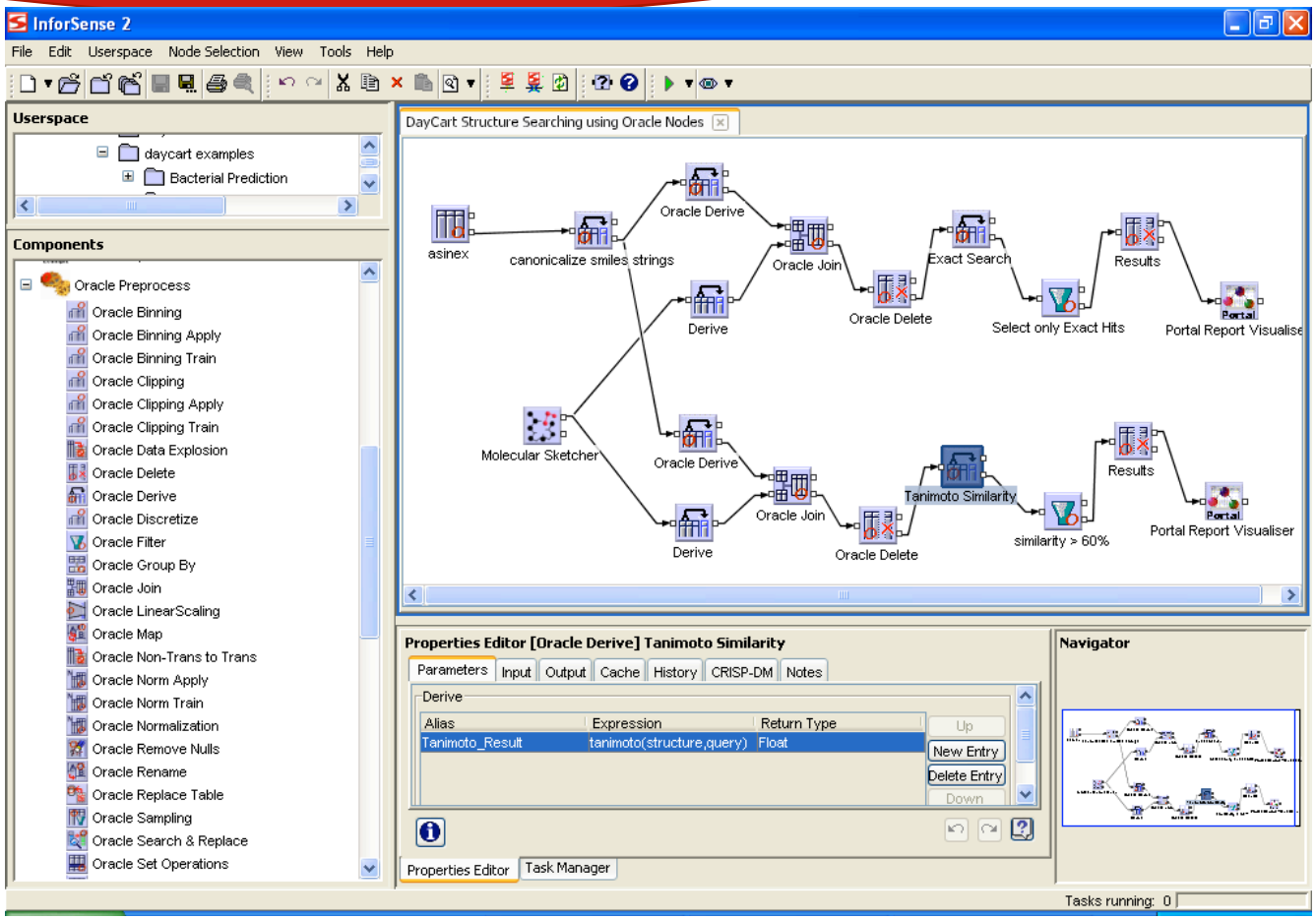
- **Generate Fingerprints**
  - Operator smi2fp – User able to define min, max and number of bits
- **Fold Fingerprints**
  - Operator foldfp – User able to define number of bits and density
- **Fingerprint Statistics**
  - Operators bitcount, nbits and isfp – Returns the number of bits and total size of the fingerprint. Also, performs a fingerbit check so that the syntax of a fingerprint can never be confused with a valid SMILES string.

Chemical Searching Nodes

- **Substructure Match / Filter**
  - Operators exact, contains
  - Match searches entire database and returns 1 or 0
  - Filter restricts searches for 1 or 0 taking advantage of DayCart chemical indexing
- **Similarity Search**
  - Operators tanimoto, euclid, Tversky
- **Reaction Search**
  - Operators reactant, agent, product, component.

Working DayCart Prototype Nodes Using Existing IOE nodes

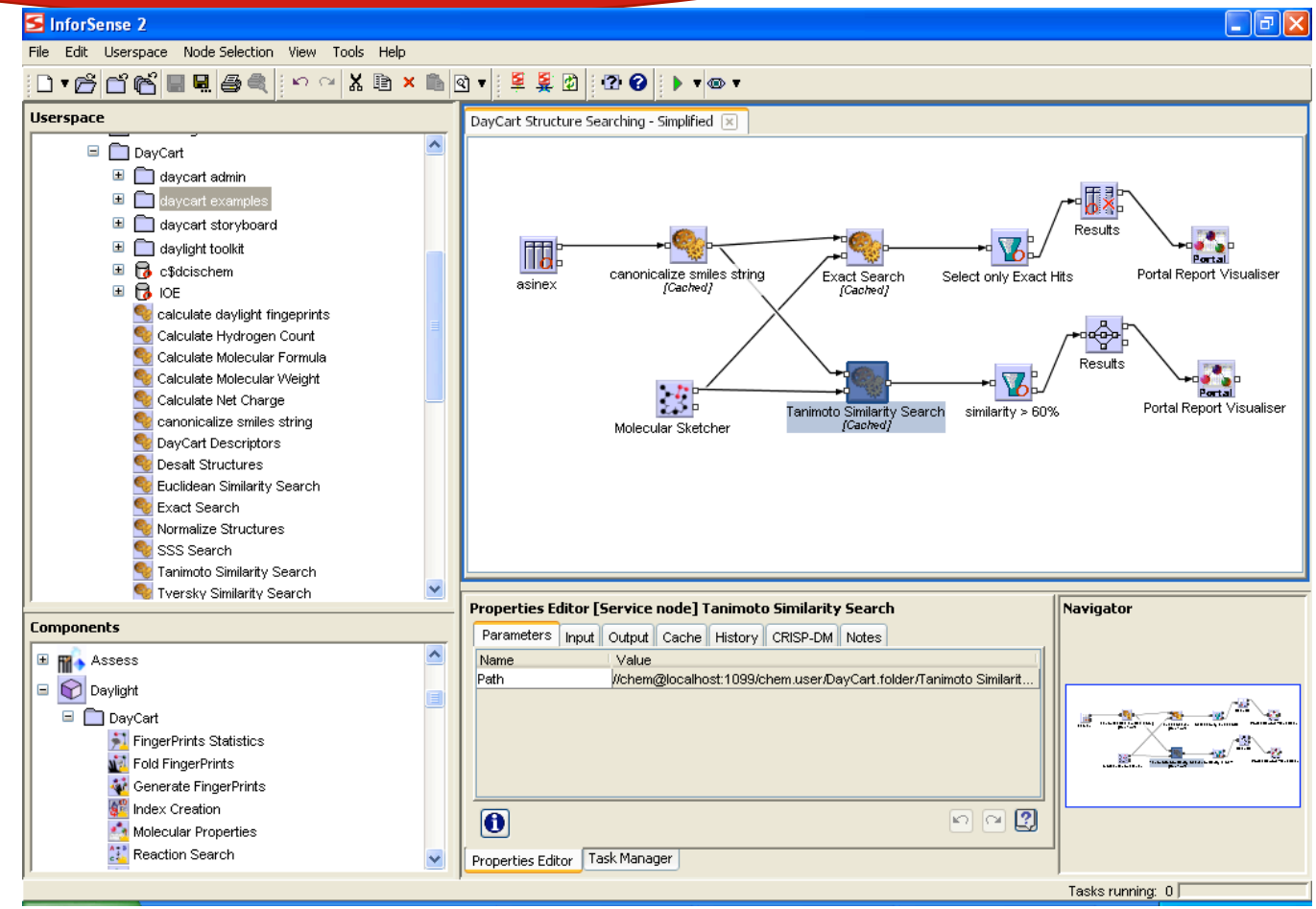
InforSense



- **Existing IOE derive or search & replace nodes can be quickly used to integrate other DayCart functions (e.g. graph, tautomer, asmiles)**

Prototype of Reusable DayCart Service Nodes

InforSense



- **Encapsulated and Reusable Group or Service nodes**
- **Installed & created prototype within 8 hours using existing oracle components**

Chemical Warehouse / Modeling Application

InforSense

- **Purpose: Create a large scale chemical warehouse consisting of commercial vendor and proprietary screening databases that is easily searched and models deployed to bench scientists architected so that all processes are performed in the Oracle databases .**
- **Six major vendor databases (Chembridge, ChemDiv, Enamine, Maybridge, Specs, Interbioscreen) were downloaded in smiles format from the Zinc web site.**
- **DayCart & Oracle Data Mining (ODM) running in Oracle 10G environment**
- **Chemical databases downloaded from http://blaster.docking.org/zinc/**
  - Ref: Irwin and Shoichet, *J. Chem. Inf. Comput. Sci. 2005;45(1):177-82*

Enamine ChemDiv CHEMICAL DIVERSITY IBS specs chemistry solutions for drug discovery CHEMBRIDGE MAYBRIDGE

Demo - Chemical Warehouse / Modeling Application

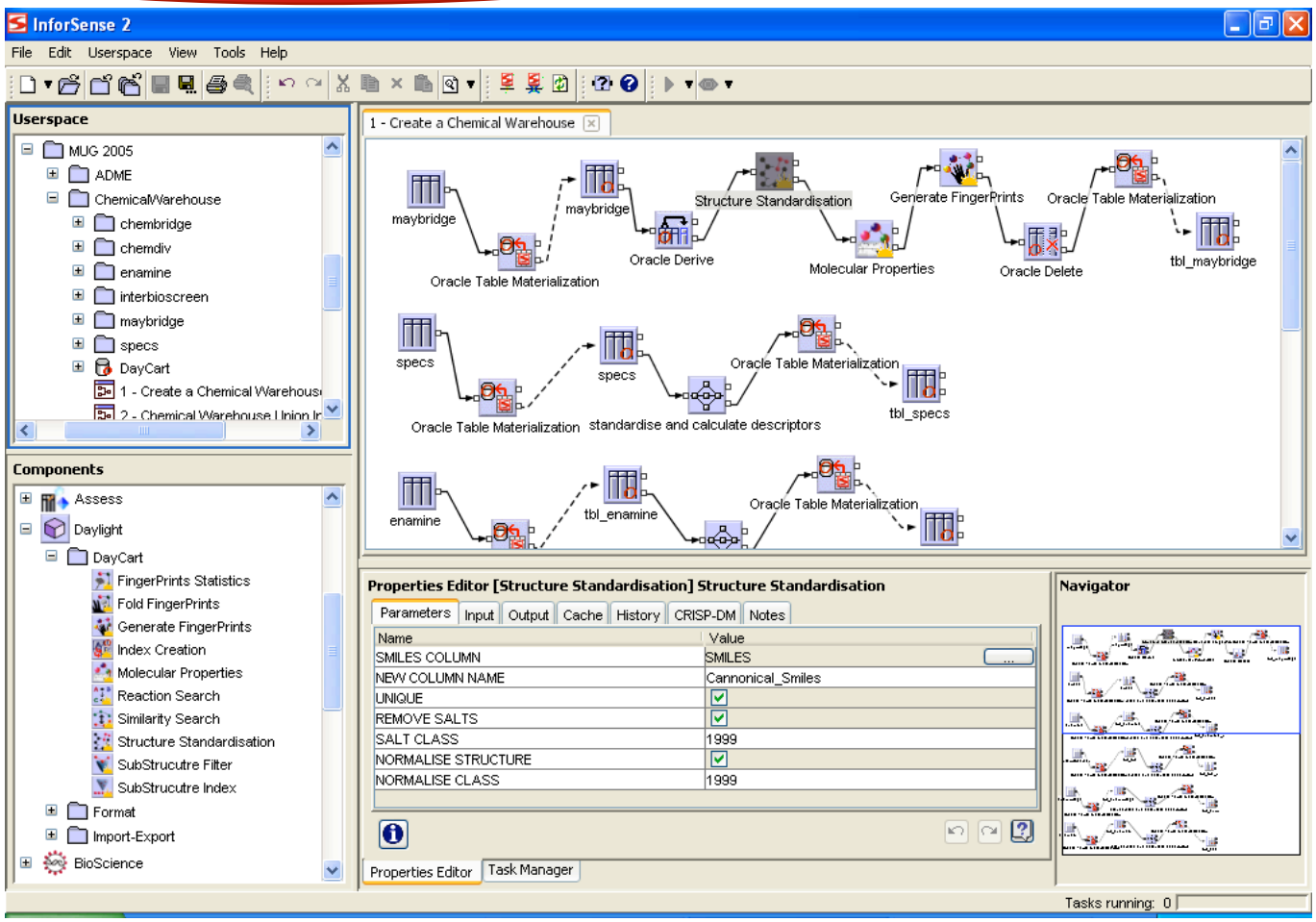
InforSense

- **Requirements - warehouse**
  - Capability of integrating with other in-house chemical databases (e.g. compound repository, subscription databases, virtual collections)
  - Easily create / modify using a visual programming environment
  - Capability of performing external or in-database data-mining analytics
  - Support chemical registrations
  - Capability of integrating with other databases (e.g. MDL, Activity Base)

- **Requirements - model**
  - ADME predictions
  - Deployment of application to benefit bench scientists
  - Oracle Data Mining Support Vector Machine (SVM) model
  - Decision Tree Classification models
  - MOE descriptors and/or Daylight Fingerprints
  - Blood Brain Barrier Permeation(1670 compounds classified BBB+ or BBB-)
  - P-glycoprotein (PgP), Human Intestinal Absorption (HIA), Torsades de Pointes (TdP)

Data Entry and Processing of Commercial Databases using DayCart

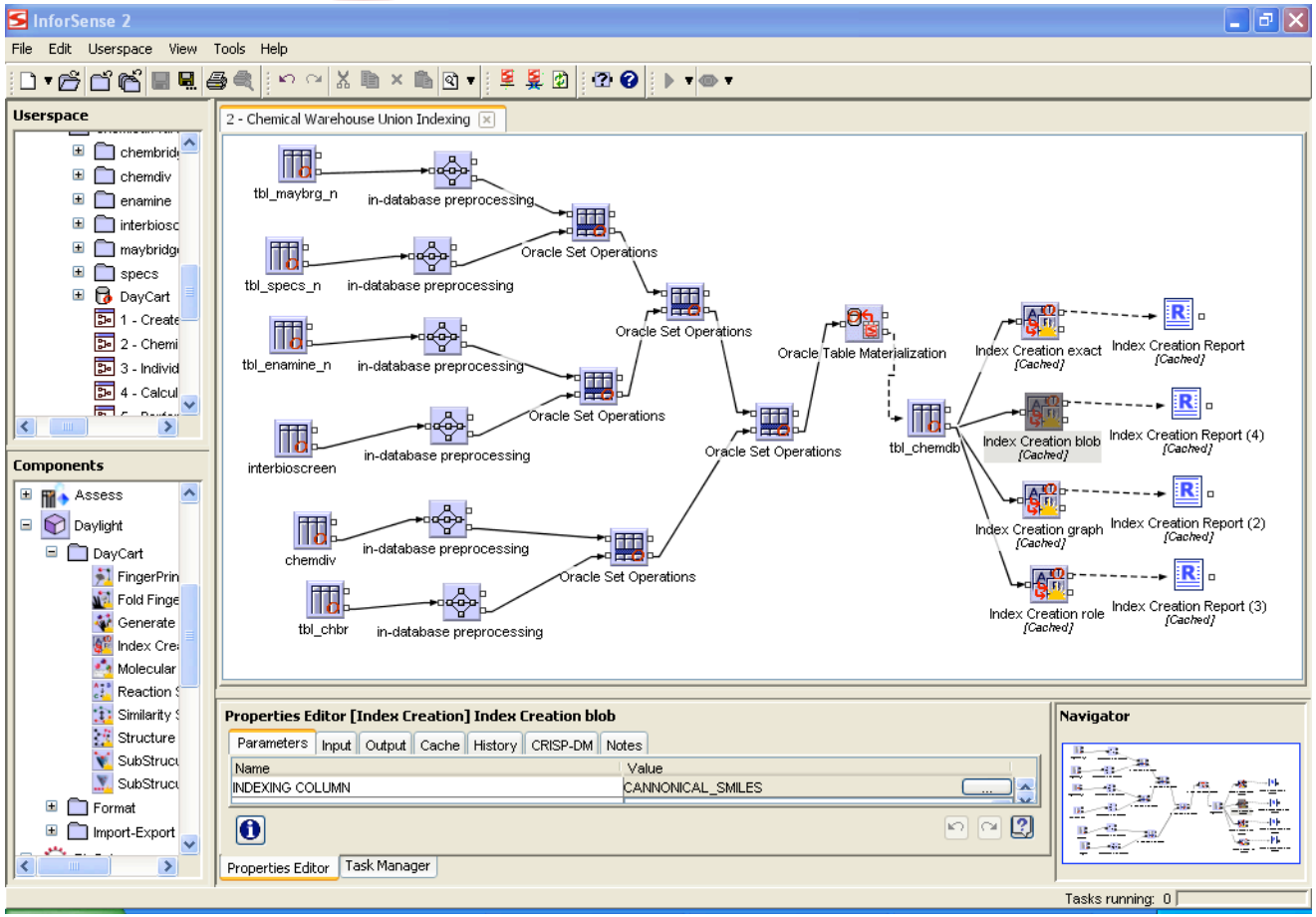
InforSense



Description: Individual commercial compound vendor database (e.g. Maybridge) are entered into an Oracle database. Structure Standardization, molecular properties and fingerprint operations are being performed using Daycart to create a table all within the Oracle Database

Creating a chemical warehouse via in-database processing

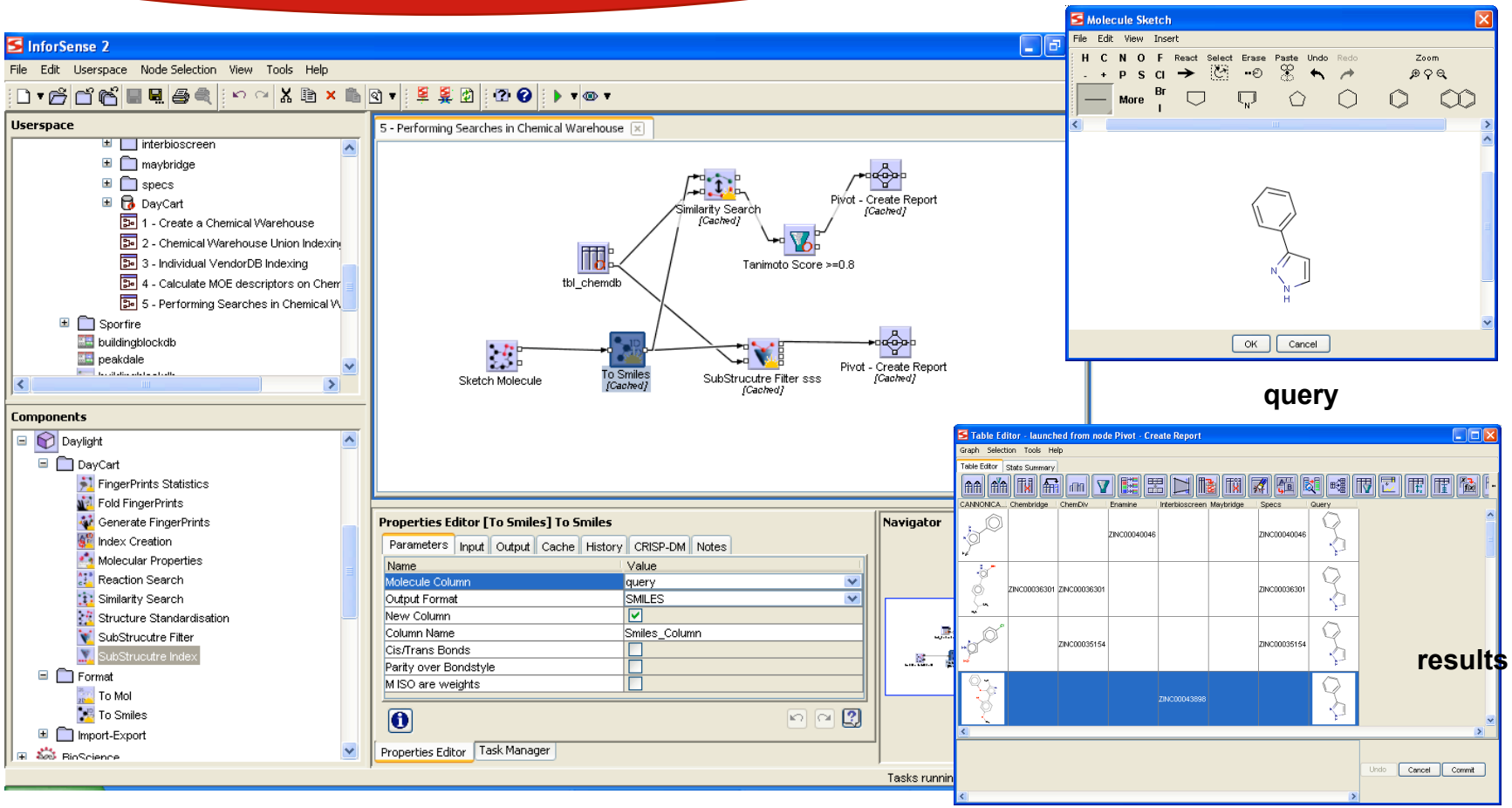
InforSense



Description: Six commercial 'cleaned' compound vendor databases are being preprocessed to unionize matching data columns (e.g. structure, CAS number, name, availability) creating a master chemical data warehouse within Oracle. DayCart structure Indexing (exact, role, graph, blob) was performed to enable fast structure searching on over over 1 million compounds.

DayCart Chemical Substructure and Similarity Searching

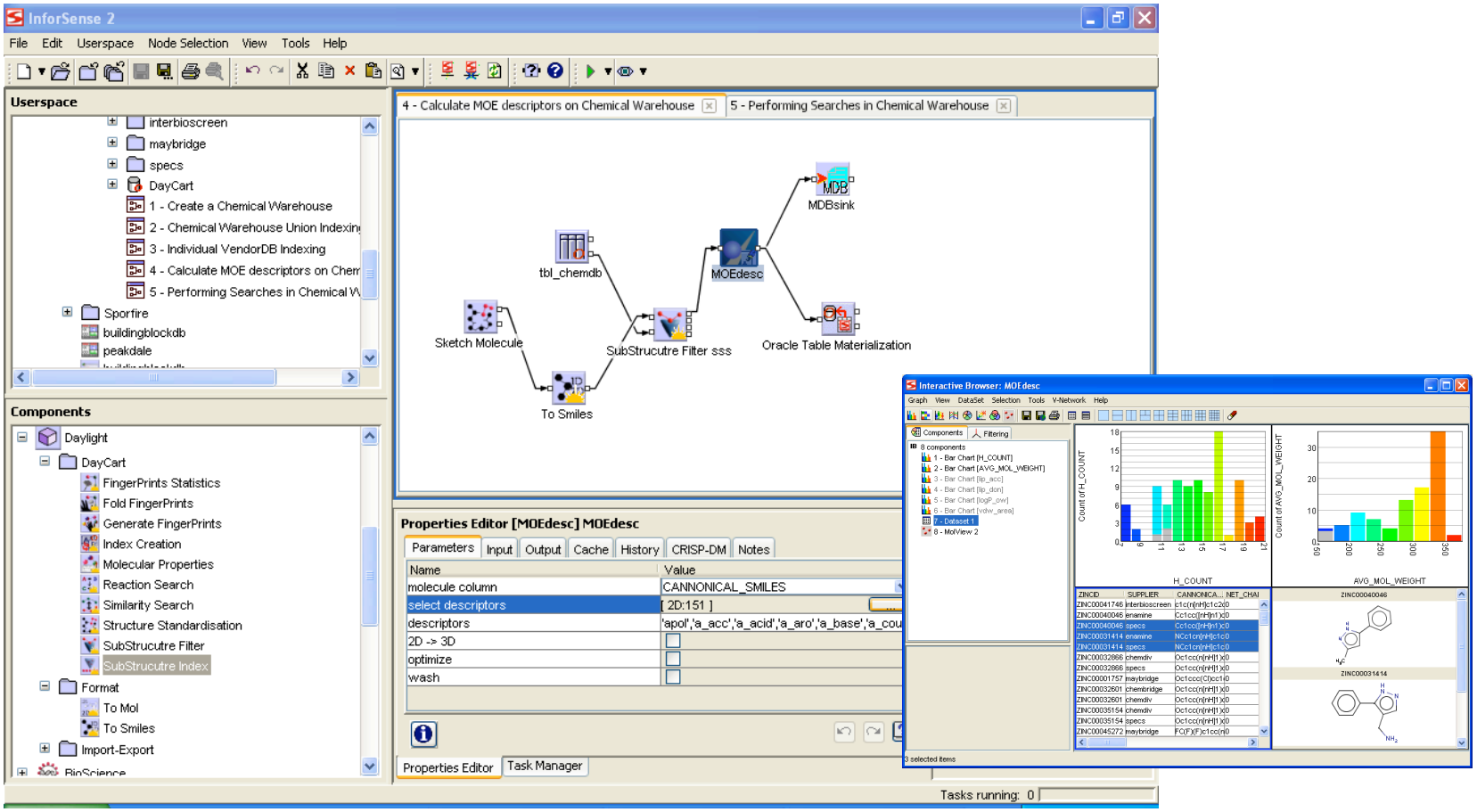
InforSense



Description: A similarity or substructure search is being performed using DayCart. Results are filtered and pivoted to produce a report of structure, query, and vendor with catalog numbers for ordering.

Calculate Molecular Descriptors using MOE

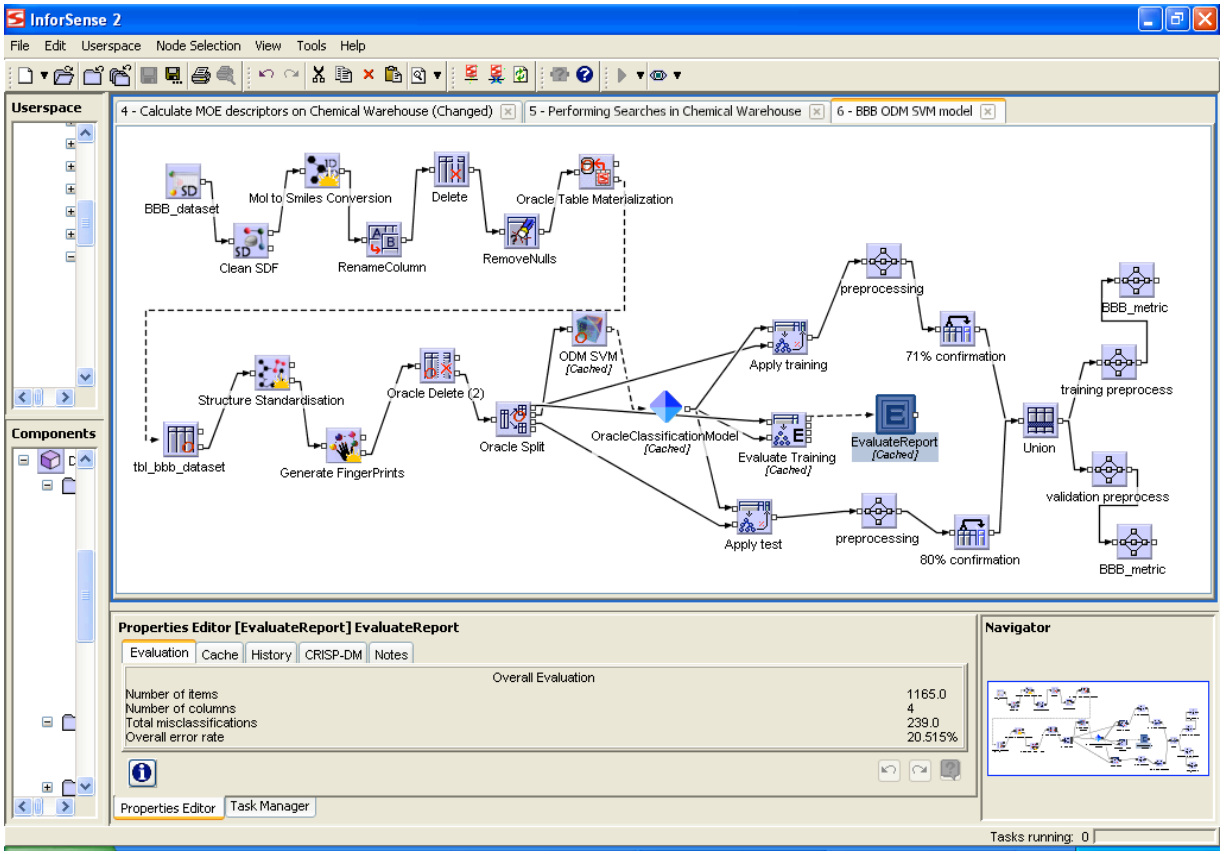
InforSense



Description: MOE descriptors (2D) are being calculated on a selection from the chemical warehouse. Results can either be used to create a MOE database or materialized into an Oracle table. Calculated molecular descriptors can be graphically viewed in an interactive browser.

KDE / IOE Blood Brain Barrier Model

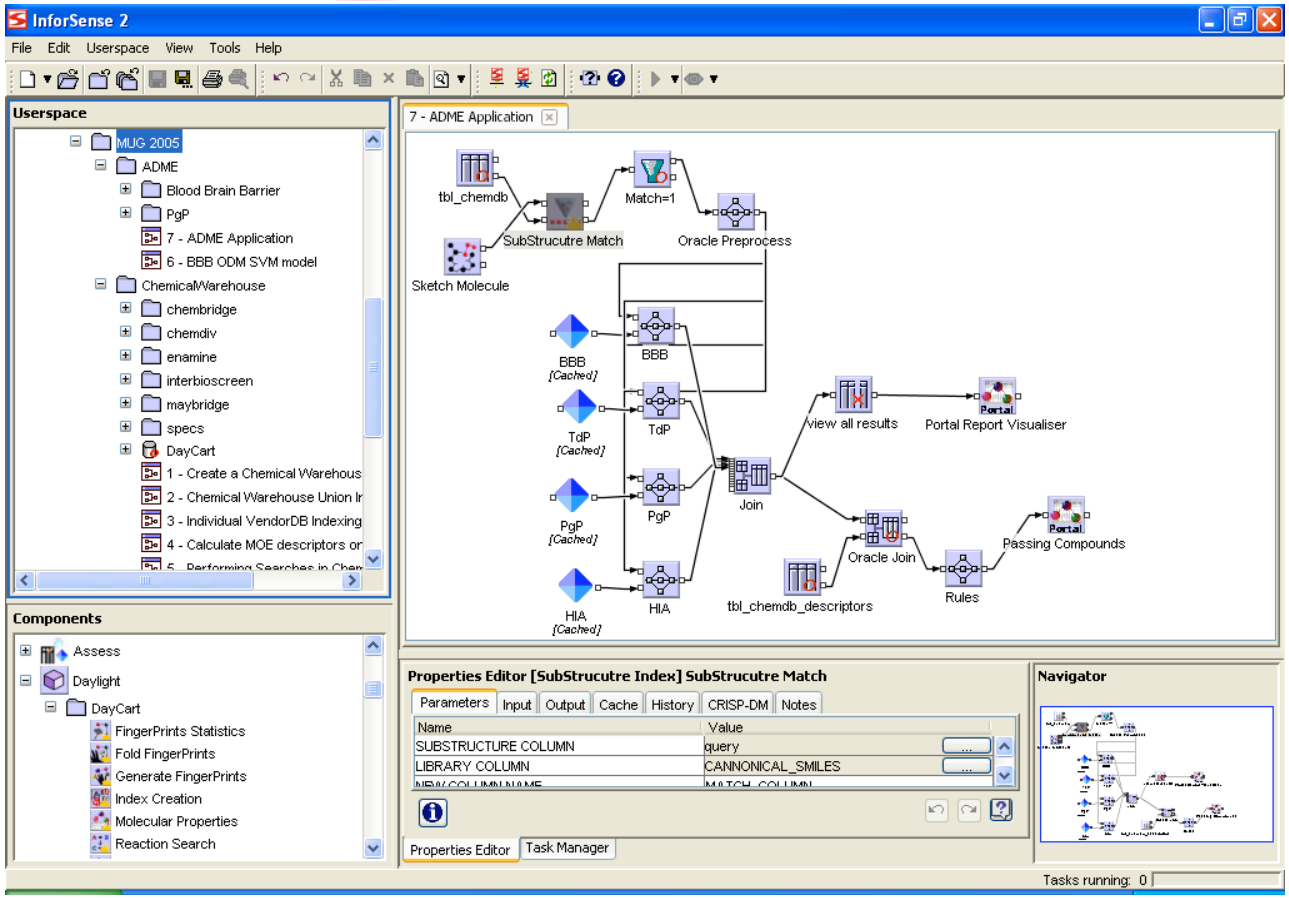
InforSense



Description: A Blood Brain Barrier dataset in SD format is converted into Daylight Smiles via 'contrib' code and imported into an Oracle table. The dataset is standardized, daylight fingerprints calculated and used to produce an SVM model all within the Oracle database. Results are then processed to calculate evaluation metrics to determine the efficiency of the model.

Modeling Application - ADME

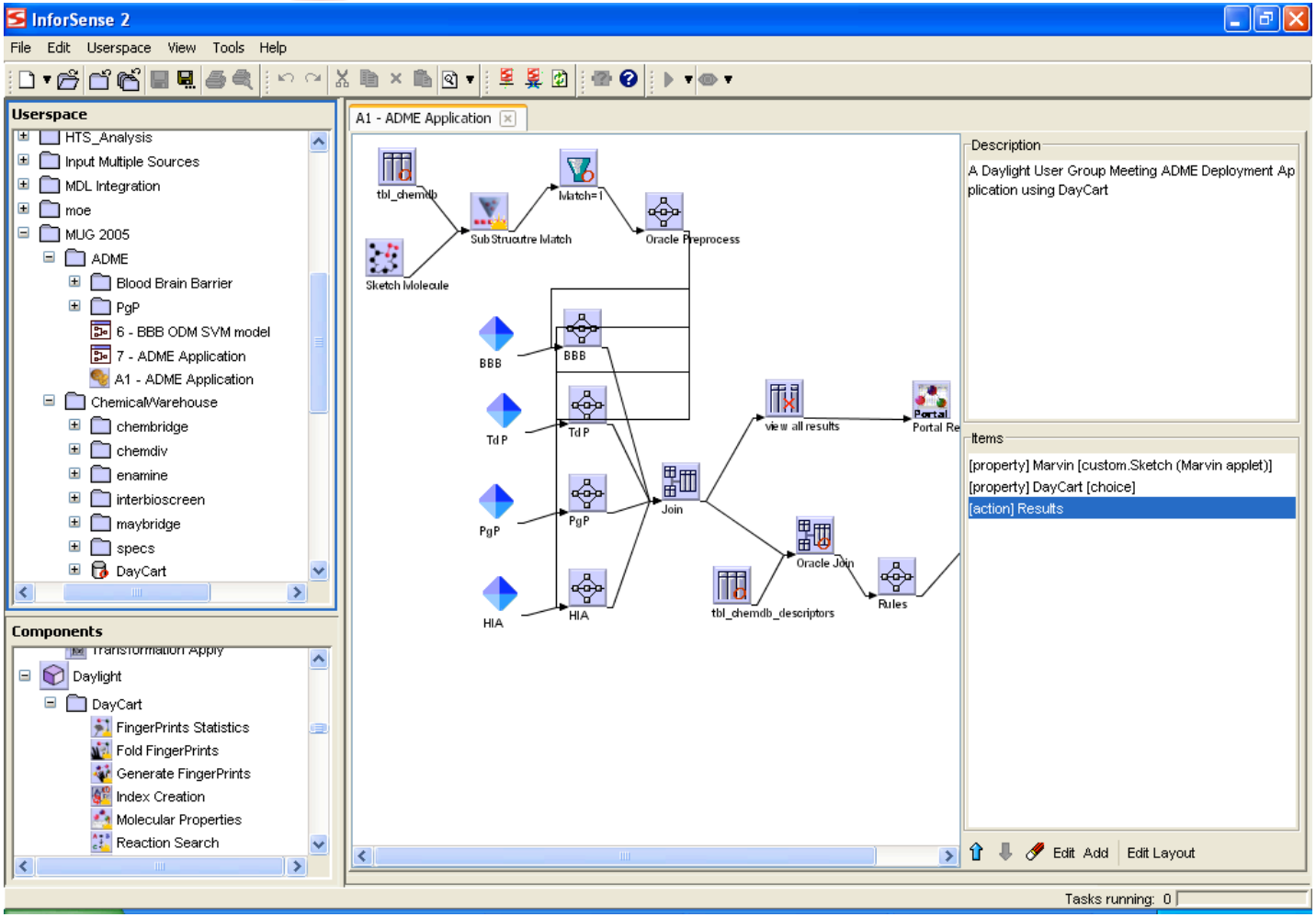
InforSense



Description: Results of a chemical search, performed in DayCart, are used to predict Blood Brain Barrier (oracle SVM), P-glycoprotein (Decision Tree), human intestinal absorption (Weka j48 DT classifier), Torsades de Pointes (oracle SVM). Predicted results are deployed via a web portal.

Deployed Application to Bench Scientists

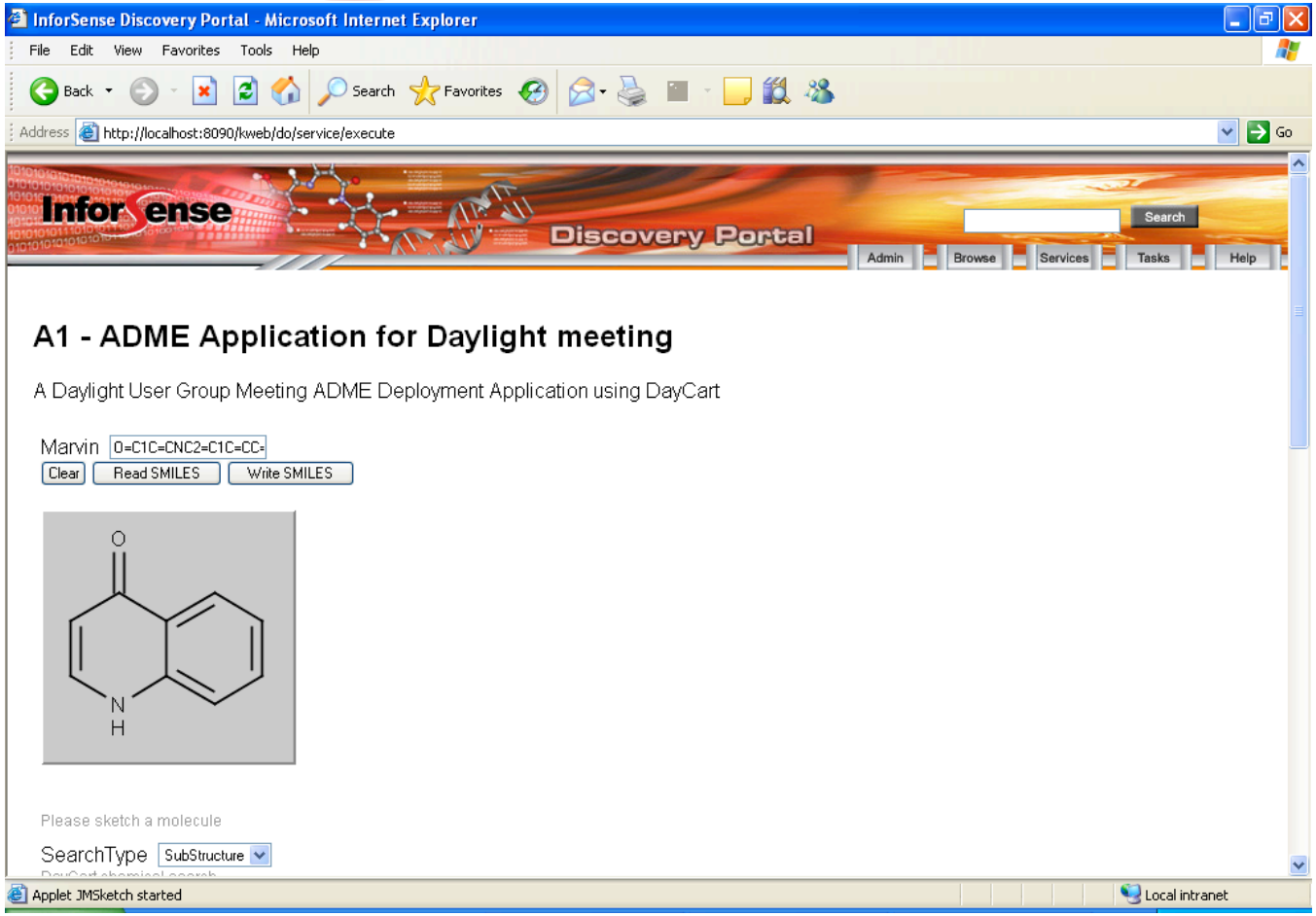
InforSense



Description: The predictive data mining ADME suite is deployed to scientists within Inforsense KDE.

Deployed Application to Bench Scientists

InforSense

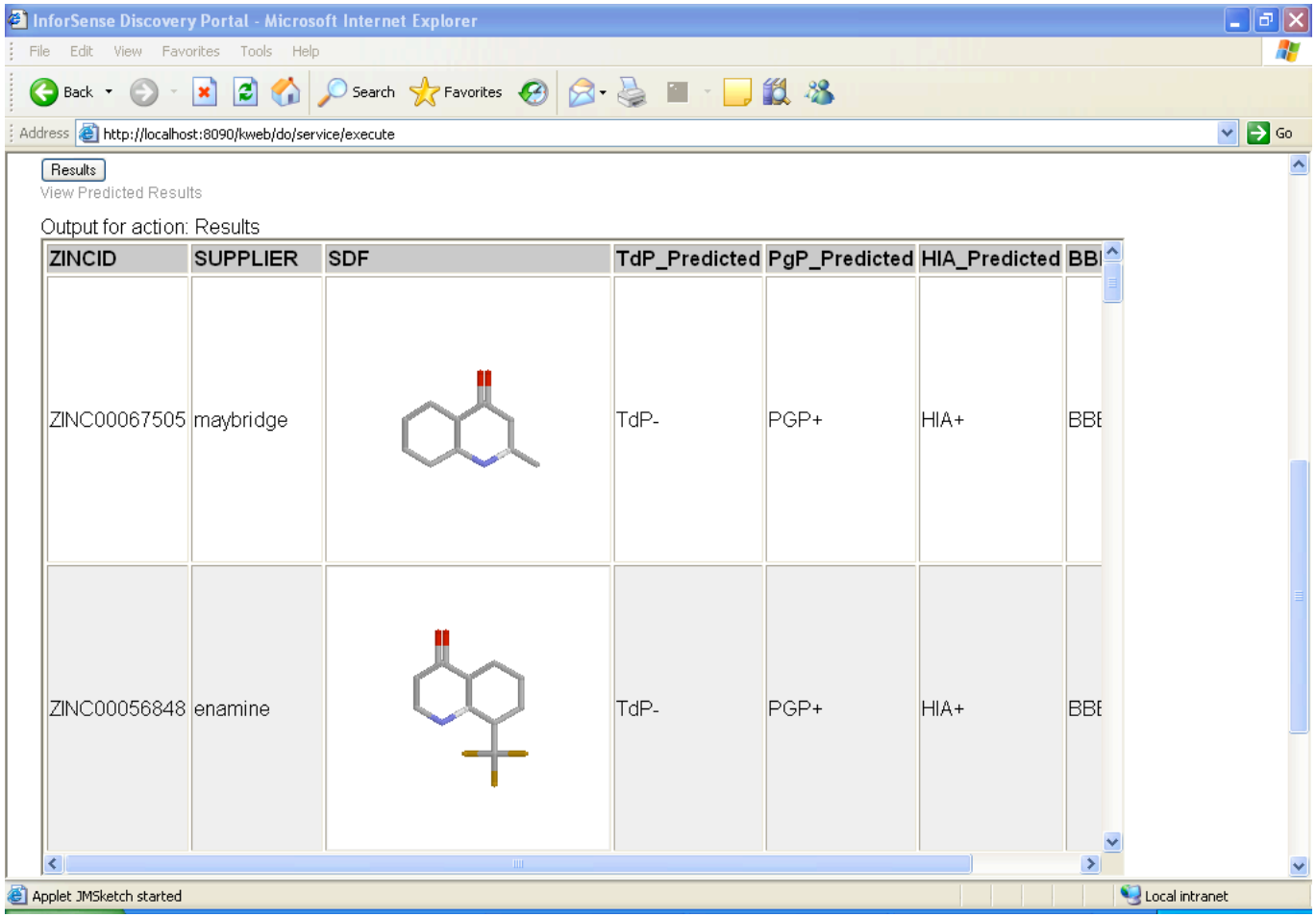


Description: The deployed predictive ADME application is presented via a web portal. Scientists can sketch a molecule using Marvin, select the type of search to be performed using DayCart, and results will be presented in tabular format.

Continued on next slide

Deployed Application to Bench Scientists

InforSense

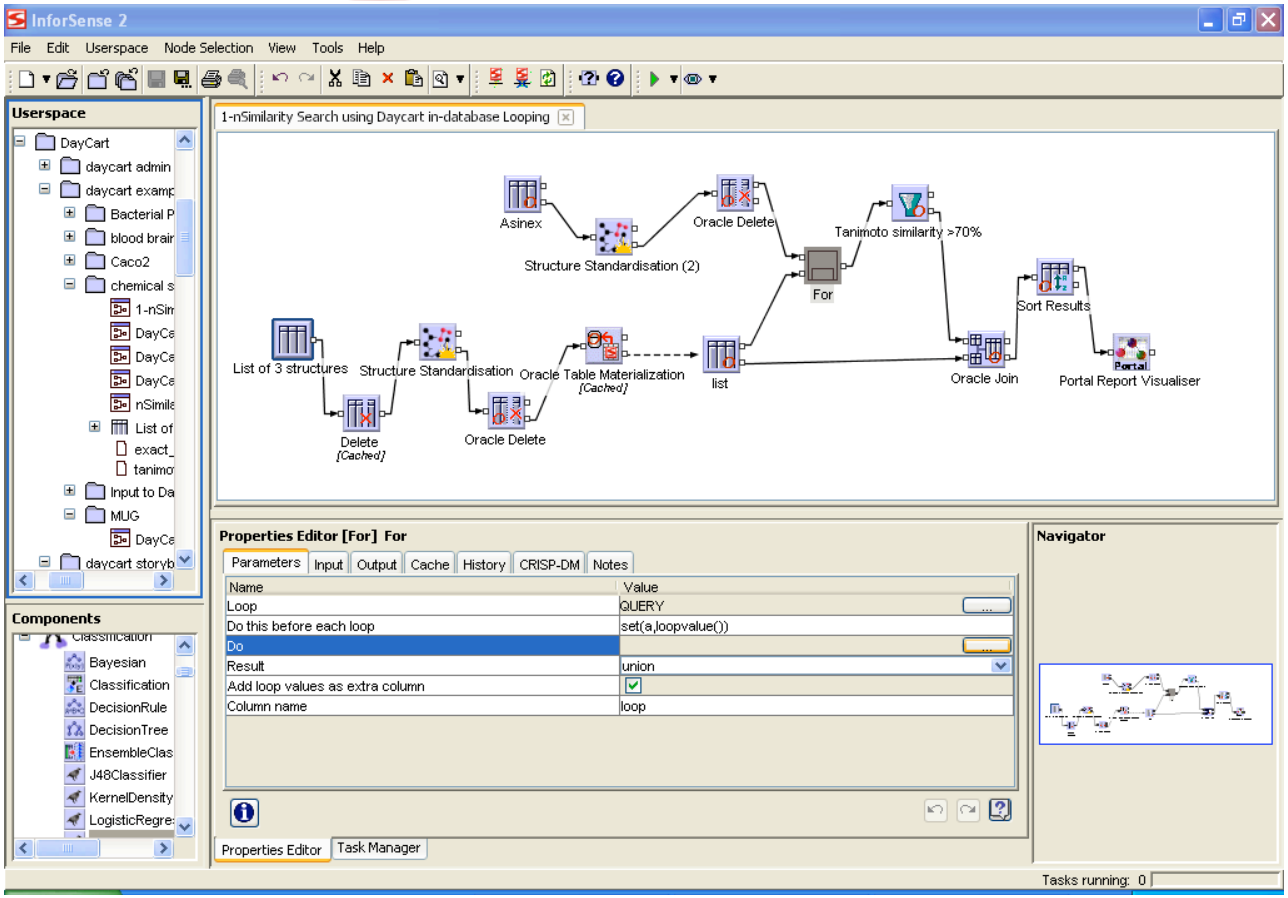


ZINCID	SUPPLIER	SDF	TdP_Predicted	PgP_Predicted	HIA_Predicted	BBI
ZINC00067505	maybridge		TdP-	PGP+	HIA+	BBI
ZINC00056848	enamine		TdP-	PGP+	HIA+	BBI

Description: Chemical structures are shown with predicted ADME results (TdP, PgP, HIA and BBB). The chemical structures are visualized using MDL Chime.

Multiple in-database chemical searching using Control (Looping) Functionality

InforSense



Description: A list of chemical structures are standardized and used as queries to be performed against a vendor database. The For control looping mechanism search each query individually and filters out all compounds with a Tanimoto similarity of less than 70 %. These processes occur all within the Oracle database.

Model References

InforSense

- J. Chem. Inf. Comput. Sci., Vol. 44, pg 1630-1638 (2004) Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents.
- J. Chem. Inf. Comput. Sci., Vol. 44, No.4 (2004), Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach.
- J. Chem. Inf. Comput. Sci.,Vol. 38, 726-235 (1998) Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure.
- J. Chem. Inf. Comput. Sci., Vol. 44, No.1 (2004) Blood-Brain Barrier Permeation Models: Discrimination between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates.

Contact InforSense

InforSense

InforSense Limited U.K
48 Princes Gardens
London SW7 2PE (UK)

Telephone: +44 (0) 20 7594 6817
Fax: +44 (0) 20 7594 6836

InforSense Limited U.S.A
25 Moulton Street
Cambridge, MA 02138 (USA)

Telephone: +1 617 547 2500

Ryoka Systems Inc.
1-28-38 Shinkawa, Chuo-Ku, Tokyo
104-0033 (Japan)

Telephone: 03-3553-6341

Additional information can be found at

www.inforsense.com