Oracle Data Mining for Real-time Analytics

NYOUG Sep 21, 2006

Shyam Varan Nath Oracle Corporation



Agenda

- Introduction to Data Mining
- What is Real-time Analytics?
- Overview of Oracle Data Mining (ODM)
- Real-time Analytical Applications
- ODM demo

About me...

- Principal Analytical Consultant, Business
 Intelligence and Warehousing group at Oracle
- Oracle Certified Professional, DBA track, since 1998 – Oracle database version 7.3 onwards
- Worked in industry domain like Finance industry, Telecomm, Healthcare, Crime Detection project etc.
- Speaker in Oracle Open world (2003), IOUG/Collaborate (2005-06), NYOUG (June), IEEE conferences etc on Data Mining





OVERVIEW

Data Mining

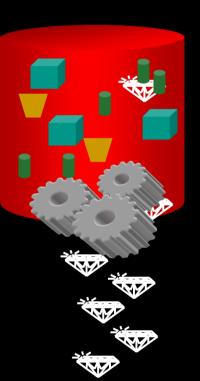
What is Data Mining?

Simply stated Data Mining refers to extracting or "mining" knowledge from large amount of data. The term is actually a misnomer, data mining should be more appropriately named "knowledge mining from the data". Thus such a misnomer that carries both "data" and "mining" became popular choice. The larger meaning encapsulates a variety of techniques and methodologies that among other things include 'cluster analysis', 'classification', 'association rules': pattern recognition in the data.

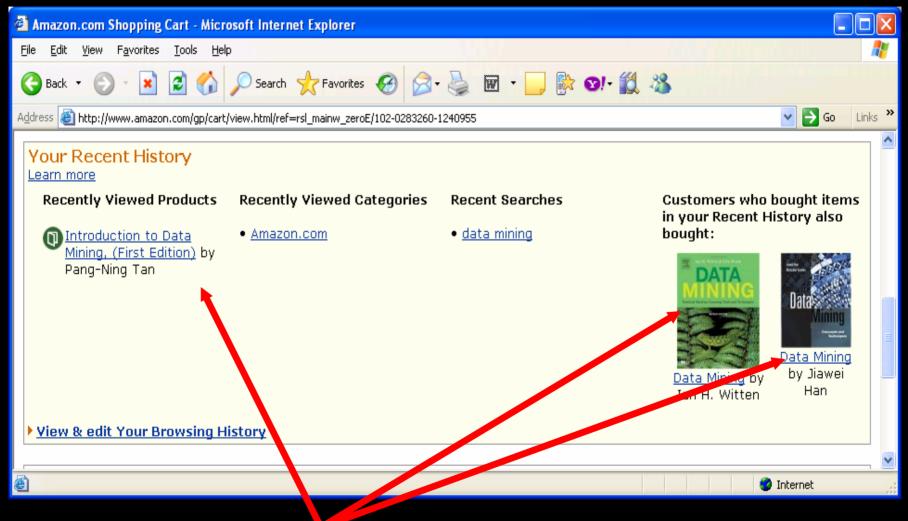
- J. Han and M. Kamber

In other words...

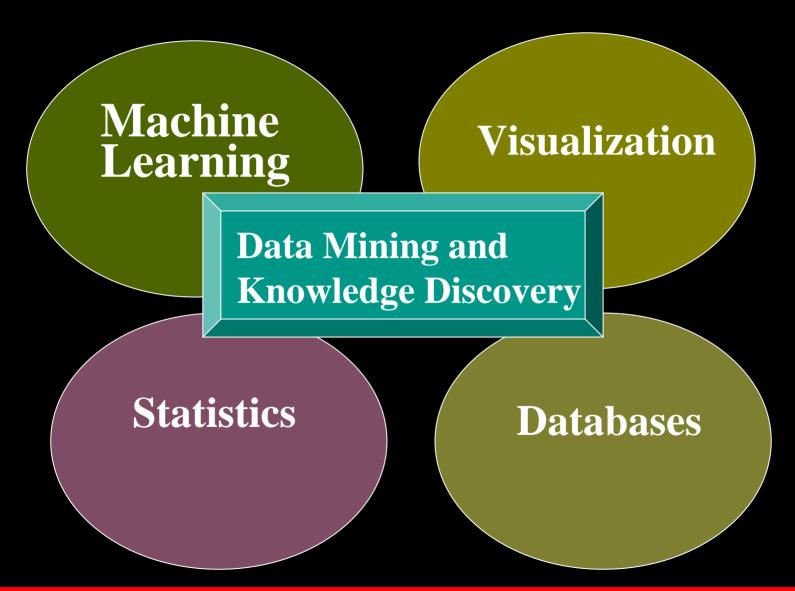
- Process of sifting through massive amounts of data to find *hidden* patterns and discover new insights
- Data Mining can provide valuable results:
 - Identify factors more associated with a target attribute (Attribute Importance)
 - Predict individual behavior (Classification)
 - Find profiles of targeted people or items (Decision Trees)
 - Segment a population (Clustering)
 - Determine important relationships with the population (Associations)
 - Find fraud or rare "events" (Anomaly Detection)



The Amazon Example (Association Rules)



Related Fields



Statistics, Machine Learning and Data Mining

- Statistics:
 - more theory-based
 - more focused on testing hypotheses
- Machine learning
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics areas not part of data mining
- Data Mining and Knowledge Discovery
 - integrates theory and heuristics
 - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

Business Intelligence

Query and Reporting

OLAP

Data Mining

Extraction of detailed and roll up data

"Information"

How is the interest rate changing in last 2 years?

Summaries, trends and forecasts

"Analysis"

What is the average home price of the condos, by region, by year of construction?

Knowledge discovery of hidden patterns

"Insights & Prediction"

Who will reconsolidate loans in the next 6 months and why?

Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

Data mining is not

- Brute-force crunching of bulk data
- "Blind" application of algorithms
- Going to find relationships where none exist
- Presenting data in different ways
- A database intensive task
- A difficult to understand technology requiring an advanced degree in computer science



Data Mining Is



- A hot buzzword for a class of techniques that find patterns in data
- A user-centric, interactive process which leverages analysis technologies and computing power
- A group of techniques that find relationships that have not previously been discovered
- A relatively easy task that requires knowledge of the business problem/subject matter expertise

Other Industry Examples

Financial Services

- Combat attrition (churn)
- Fraud detection
- Loan default
- Identify selling opportunities

Telecommunications

- Identify customers likely to leave
 Target highest lifetime value
 customers
- Identify cross-sell opportunities

Retail

- Loyalty programs
- Cross-sell
- Market-basket analysis
- Fraud detection

Database Marketing

- Buy product x
- More targeted & successful campaigns
- Identify cross-sell & up-sell opportunities

Insurance, Government

- Flag accounting anomalies (Sarbanes-Oxley)
- Reduce cost of investigating suspicious activity or false claims

Life Sciences

- Find factors associated with healthy or unhealthy patients
- Discover gene and protein targets
- Identify leads for new drugs

What is Real-time Analytics?

- Real-time analytics is the use of, or the capacity to use, all available enterprise data and resources when they are needed. It consists of <u>dynamic</u> <u>analysis</u>, drawing inferences and reporting, based on data entered into a system up to the actual time of use.
- Real-time analytics is also known as real-time data analytics, real-time data integration, and realtime business intelligence.

Examples of Real Time Analytics

- Real-time analytics can be used in <u>CRM</u>
 (customer relations management) analytics, which
 includes all programming that analyzes data about
 an enterprise's customers and presents it so that
 better and quicker business decisions can be
 made.
- Another application is in scientific analysis such as the tracking of a hurricane's path, intensity, and wind field, with the intent of predicting these parameters hours or days in advance.

Data Mining Techniques (10g, 10gR2)

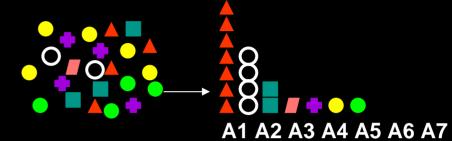
- Unsupervised (clustering)
- K-means
- O-Cluster
- Non-Negative Matrix Factorization
- Anomaly Detection
- Hierarchical
- Self Organizing maps
- Expectation Maximization
- Many More (pca,mds)

- Supervised (classification and prediction)
- Adaptive Bayes Network (ABN)
- Naïve Bayes
- Support Vector Machine
- Decision Trees
- K Nearest Neighbor
- Neural Networks
- Linear Discrimination (PL/SQL code generator)

Algorithms & Example Applications

Attribute Importance

- Identify most influential attributes for a target attribute
 - Factors associated with high costs, responding to an offer, etc.

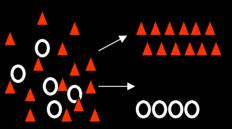


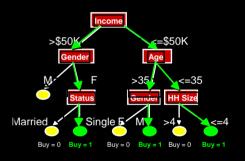
Classification and Prediction

- Predict customers most likely to:
 - Respond to a campaign or offer
 - Incur the highest costs
- Target your best customers
- Develop customer profiles

Regression

- Predict a numeric value
 - Predict a purchase amount or cost
 - Predict the value of a home







Algorithms & Example Applications

Clustering

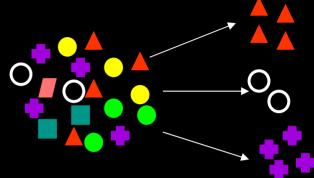
- Find naturally occurring groups
 - Market segmentation
 - Find disease subgroups
 - Distinguish normal from non-normal behavior

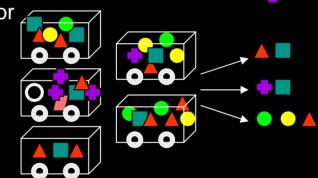
Association Rules

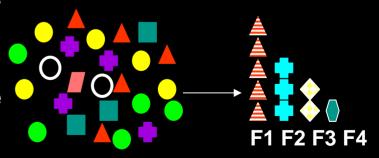
- Find co-occurring items in a market basket
 - Suggest product combinations
 - Design better item placement on shelves

Feature Extraction

- Reduce a large dataset into representative new attributes
 - Useful for clustering and text mining







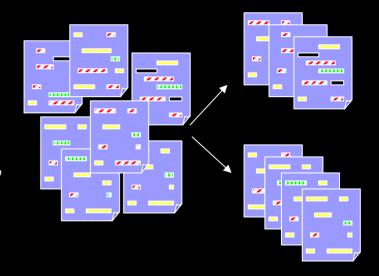
Algorithms & Example Applications

Text Mining

- Combine data and text for better models
 - Add unstructured text e.g. physician's notes to structured data e.g. age, weight, height, etc., to predict outcomes
- Classify and cluster documents
 - Combined with Oracle Text to develop advanced text mining applications e.g.
 Medline

BLAST

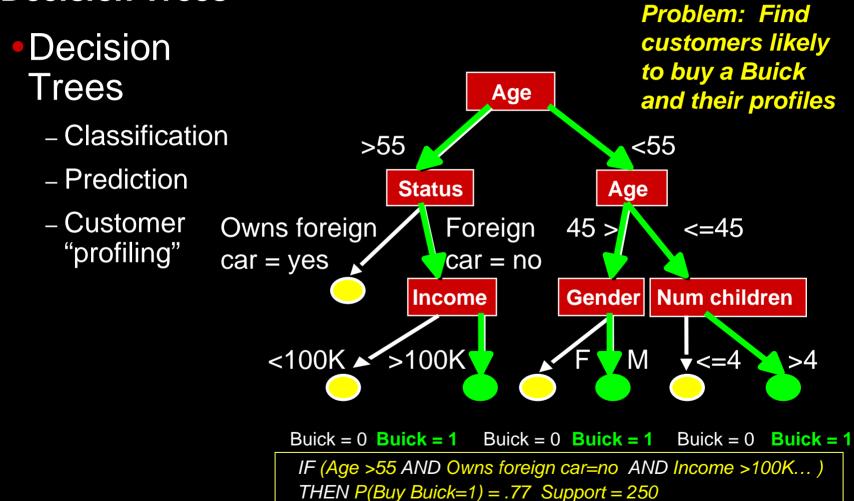
- Sequence matching and alignment
 - Find genes and proteins that are "similar"



ATGCAATGCCAGGATTTCCA

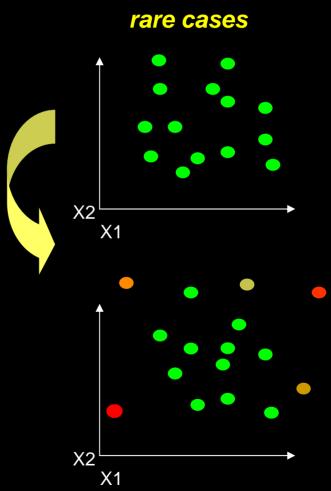
CTGCAAGGCCAGGAAGTTCCA ATGCGTTGCCAC...ATTTCCA GGC..TGCAATGCCAGGATGACCA ATGCAATGTTAGGACCTCCA

Oracle Data Mining 10g R2 Decision Trees



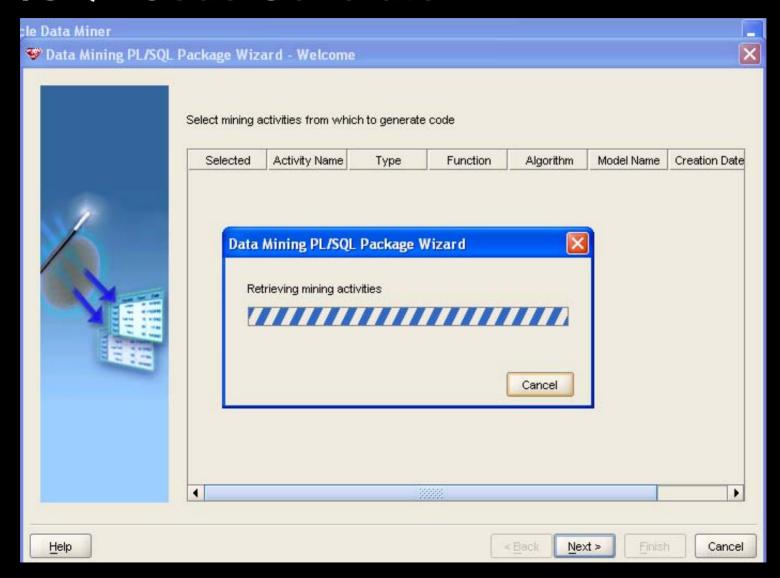
Oracle Data Mining 10g R2 Anomaly Detection

- "One-Class" SVM Models
 - Fraud, noncompliance
 - Outlier detection
 - Network intrusion detection
 - Disease outbreaks
 - Rare events, true novelty



Problem: Detect

PL/SQL Code Generator



PL/SQL code sample

```
/* pl/sql code gen1 */
```

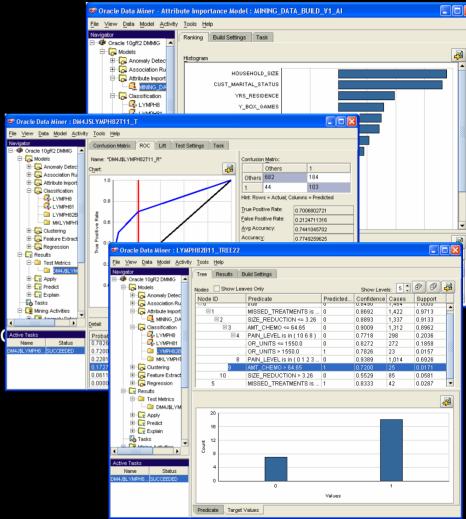
```
CREATE PACKAGE "DATAMININGACTIVITY1" AUTHID DEFINER AS PROCEDURE "MINING_BUILD_TEST" (case_table IN VARCHAR2 DEFAULT "DMUSER1". "MINING_BUILD_TEXT"), additional_table_1 IN VARCHAR2 DEFAULT NULL, model_name IN VARCHAR2 DEFAULT 'MINING_BUILD_75202_DT', test_metric_name IN VARCHAR2 DEFAULT "DM4J$MINING_TEST"', END;
/* pl/sql code gen1 */
```

CREATE PACKAGE BODY "DATAMININGACTIVITY1" AS

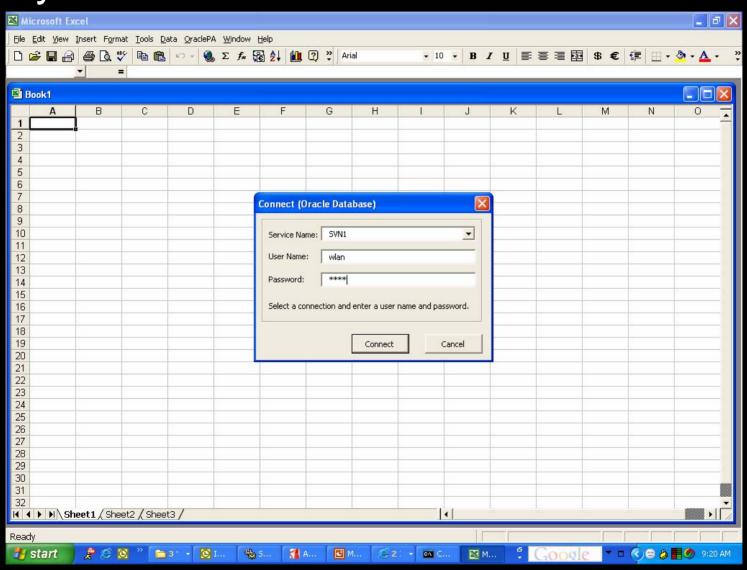
c_long_sql_statement_length CONSTANT INTEGER := 32767;

.

- Oracle mining platform
 - PL/SQL API
 - Java API (JSR-73)
 - Oracle Data Miner (gui)
 - Spreadsheet Add-In
- Range of algorithms
 - Structured & unstructured data
 - Attribute importance
 - Classification, regression & prediction
 - Anomaly detection
 - Association rules
 - Clustering
 - Nonnegative matrix factorization
 - BLAST



Oracle Spreadsheet Add-In for Predictive Analytics



Oracle Data Miner Support for Text

- ODMr allows one column of the input table for a mining activity to be a text column. You can mine tables with two or more text columns using the ODM programmatic interfaces. If you have text columns only, you can use ODM or Oracle Text.
- Data Preparation for Text Columns Any text columns must be properly prepared:
 - For the Oracle Data Mining programmatic interfaces, text columns must be converted to nested columns.
 - For an Oracle Data Miner Activity, the text column must be indexed. Oracle Data Miner automatically creates the index during the mining activity.
- Data | Transform | Text lets you prepare a text column for use with the Oracle Data Mining PL/SQL interface. The same transform allows you to do directly the processing done internally by the mining activity.

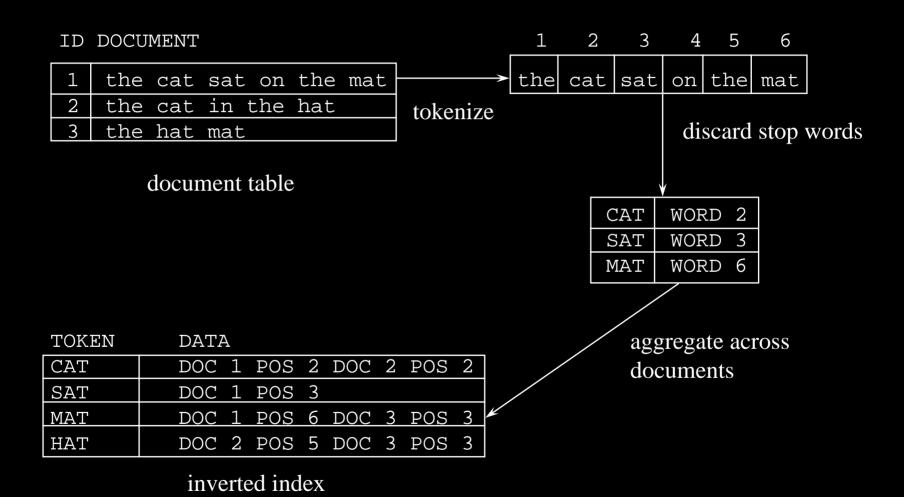
ODMr Restrictions on Text Mining

- ODMr does not support all of text mining functionality provided by Oracle Data Mining. The following restrictions apply to text mining using ODMr:
- You can include zero or one text columns in a mining operation. If you need to build a model with two or more text columns, you must use one of the ODM programmatic interfaces.
- The mining type of the text column must be text; if the data type of the column is VARCHAR2 or CHAR, you will have to change the mining type from categorical to text.
- The case ID column must be numerical.
- You must have Oracle Text installed with appropriate permissions

Indexing & PL/SQL (Oracle Text)

- Three index types
 - context for traditional information retrieval
 - ctxcat for catalogs
 - ctxrule for classification and/or routing
- Extensions to SQL
 - select ... from ... where contains ...
 - select ... from ... where catsearch ...
 - select ... from ... where matches ...
- Large set of PL/SQL packages

The Inverted Index



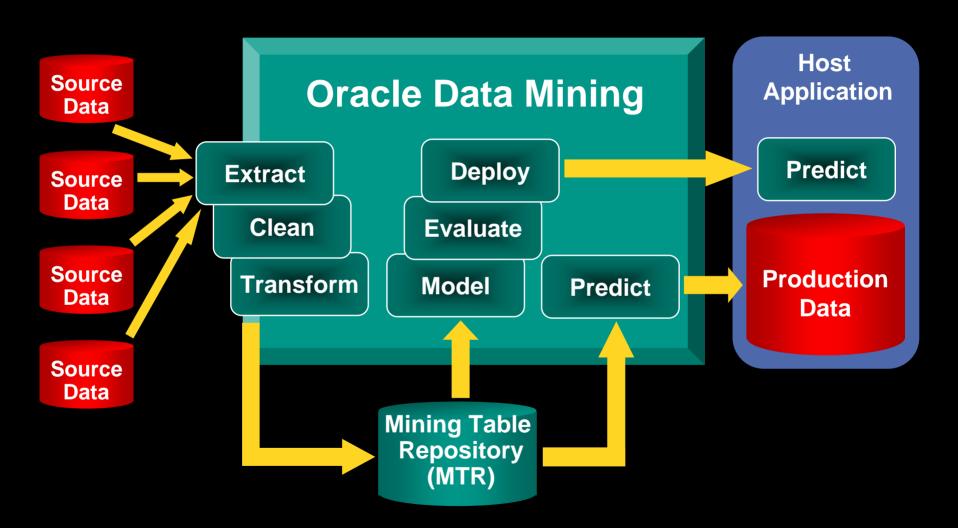
Text Retrieval

```
create table foo (text varchar2(80));
insert into foo values ('the cat sat
on the mat');
create index foox on foo(text)
indextype is ctxsys.context;
select * from foo where
contains(text, 'cat')>0;
TEXT
the cat sat on the mat
```

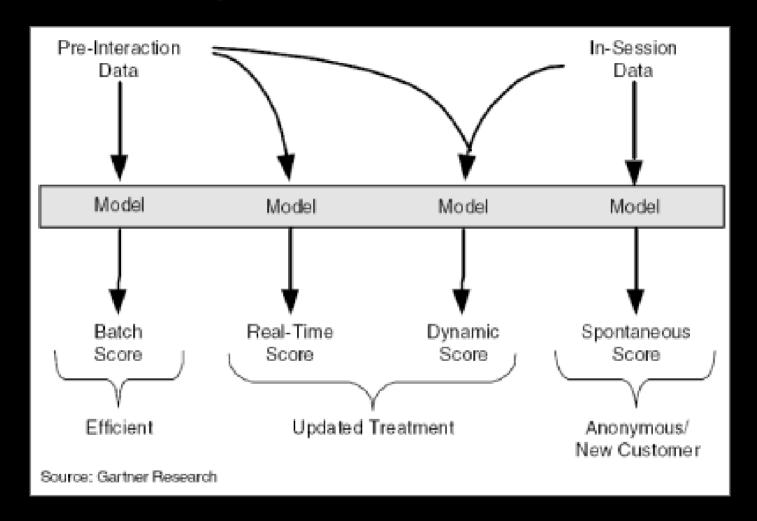
Real-Time Applications based on Data Mining

- Why do we need Real time decisions?
- Recent industry changes like "DO NOT CALL LIST", TIVO, spam blockers, aversion to junk post mail etc
- Importance of "In-Bound Opportunity"
- Finance-banking industry example

Data Mining Flow



Model Scoring Approach



Gartner's Scoring Approach Classification

Continuum of Scoring Options

Matching the approach to Business requirement

	Barrier States
Scoring Approach	Description
Batch	Most common approach. Can be scheduled, and the results can be shared among different applications. Disadvantages can be incorrect scores due to time lags, amount of time and storage needed to score numerous models with many predictors. In addition, the number of cases affects scoring.
Real Time	Scoring allows for updates on recently collected data. Especially useful when little new data is collected.
Dynamic	Scoring allows data to be integrated from current interaction. Useful if there is a significant amount of missing information that is important for the model.
Spontaneous	Used to gain insight into a new or anonymous customer for whom a previous history or profile is unavailable.







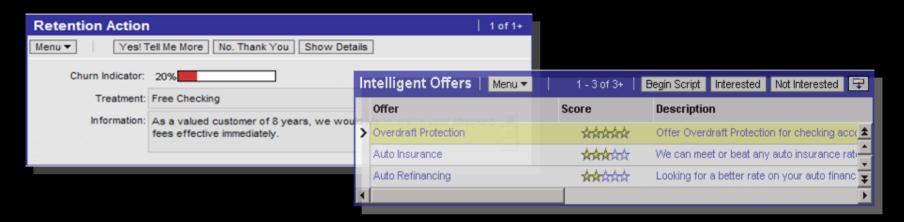
Industry Specific Cases/Applications of

Data Mining

The "in-bound" window of opportunity

- In-bound calls present a great opportunity for upsell and cross sell
- Data mining scoring for recommendations in Real time can provide CSR's with most likely loan product to offer
- Such an application was built using Oracle Data Mining for finance industry

Intelligent Offer Generation and Retention Management Application



 Integrated User Interface for real time analytical application based on data mining (from Finance Industry)

Demo Scenario: Business Challenges

About National Bank

- Fictional financial services provider
- Customer base: 5 million
- Assets: \$69 billion
- Revenue: \$4.6 billion
- Large volume Siebel Call Center

Business Challenges

- High customer turnover rate of 14% per year
- Associated replacement cost in millions per year
- Average cost of new customer acquisition: \$250
- Currently 2 products per customer, goal of achieving 4 per customer

Demo Scenario: Call Center Solution

- Predict in real-time customers propensity to attrite and to respond to various retention treatments
- Offer relevant and timely retention offers such as free online bill payment only to customers most likely to leave
- Predict in real-time customers propensity to respond to various cross- & up sell offers
- Target customers with relevant and timely cross sell offers at the time of call

instead of

Running costly, less relevant and less timely outbound retention and cross- & up-sell campaigns

Call Scenario A: Intelligent Cross-Sell

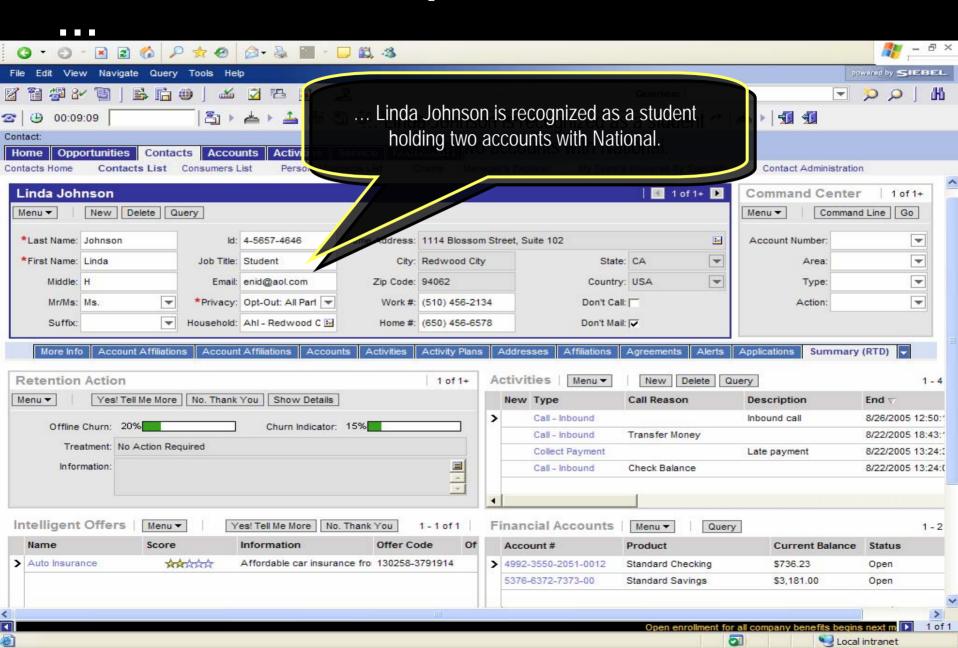
Profile of caller (Linda Johnson):

- Female, 28 years old, single
- Holds checking and savings account at National Bank
- Medium-value customer
- Calls to change address (due to new job after grad school)

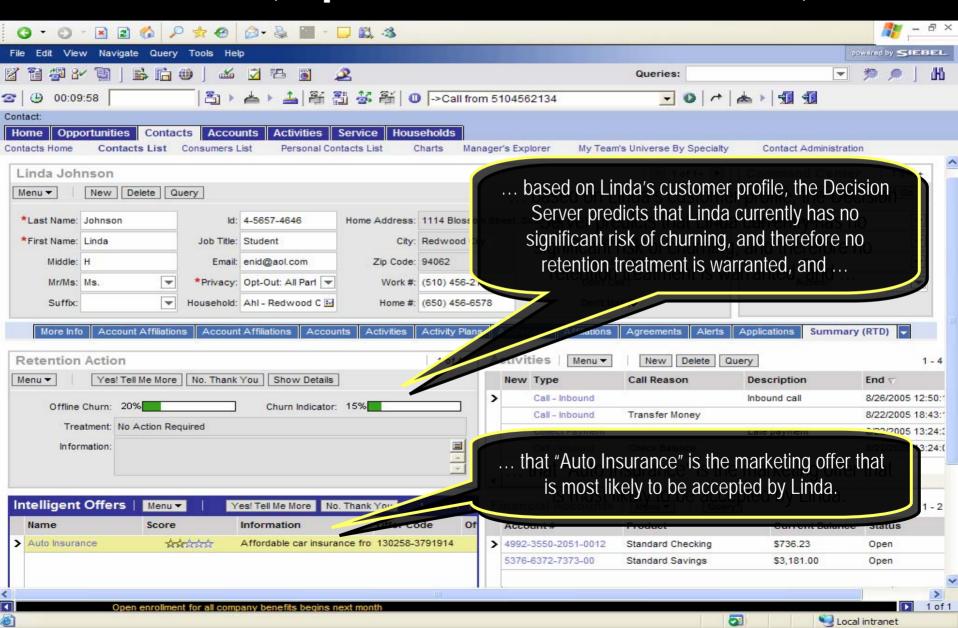
Objectives of National Bank:

Expand customer relationship through real-time intelligent cross- and up-sell offers

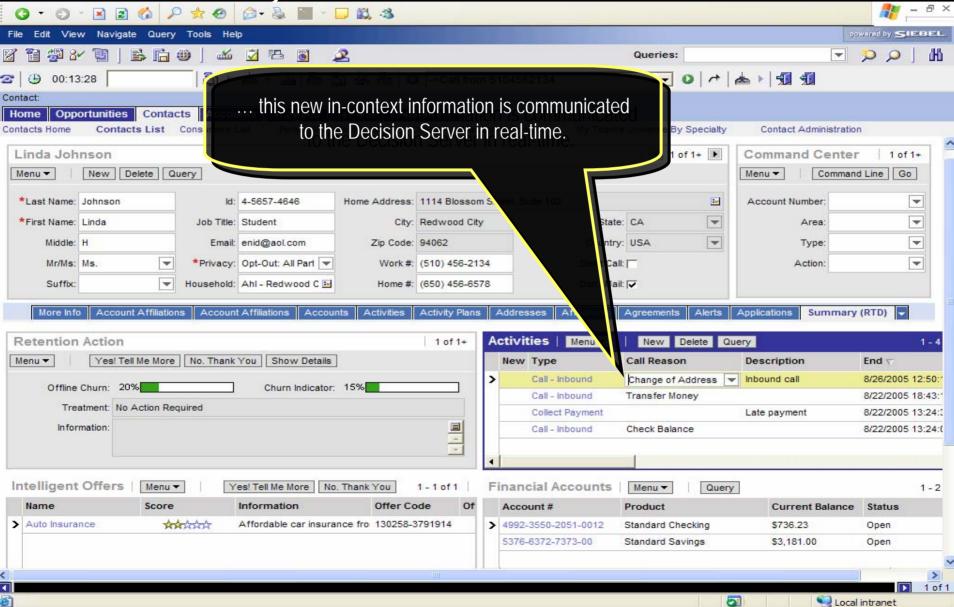
Call Scenario A: Upon caller identification,



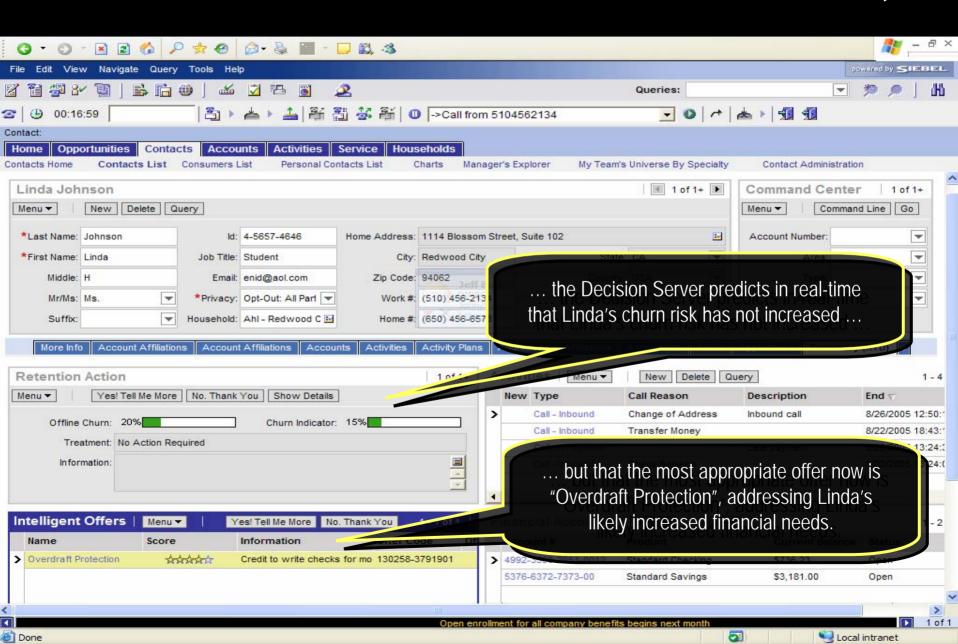
In addition, upon caller identification, ...



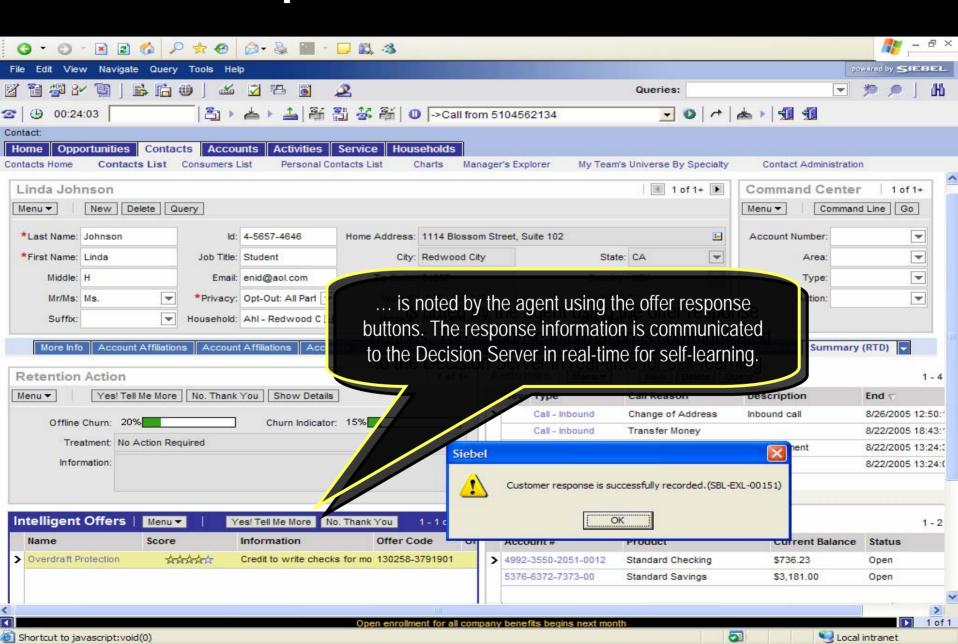
Upon noting Linda's call reason, change of address, ...



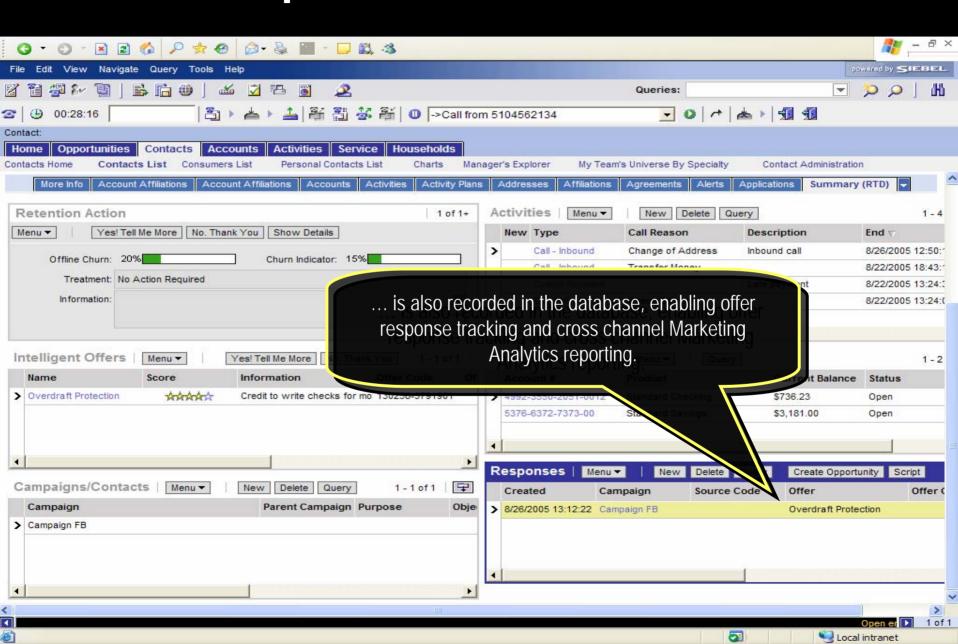
Based on the new in-context information, ...



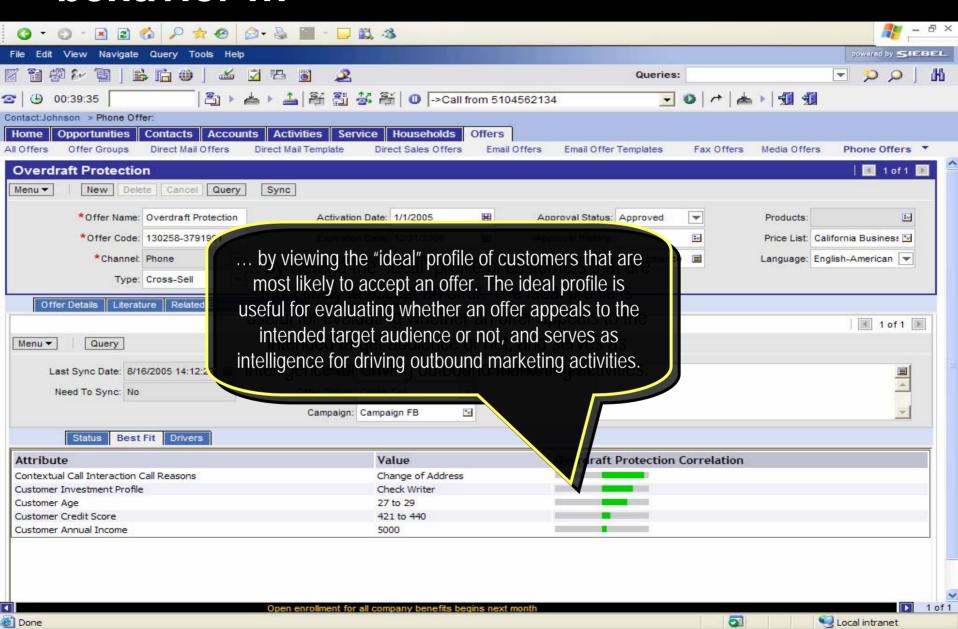
Linda's response to the extended offer ...



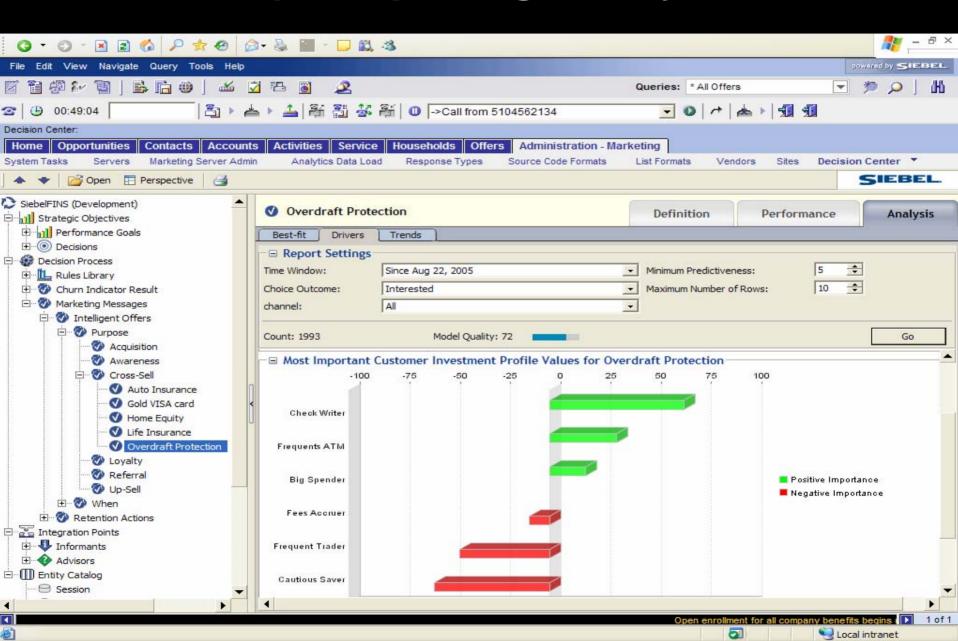
Linda's response to the extended offer ...



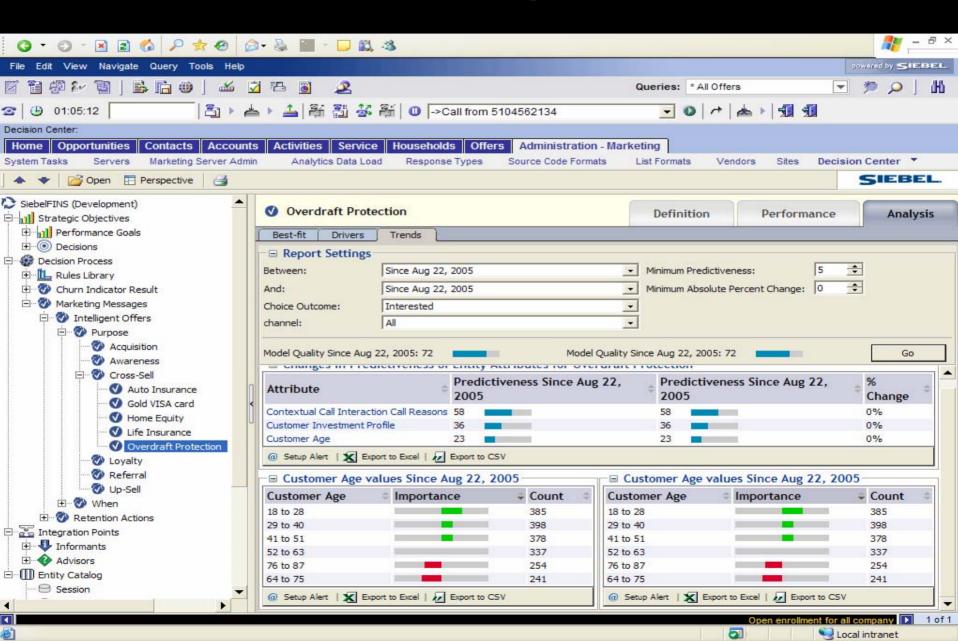
Marketing users can analyze offer response behavior ...



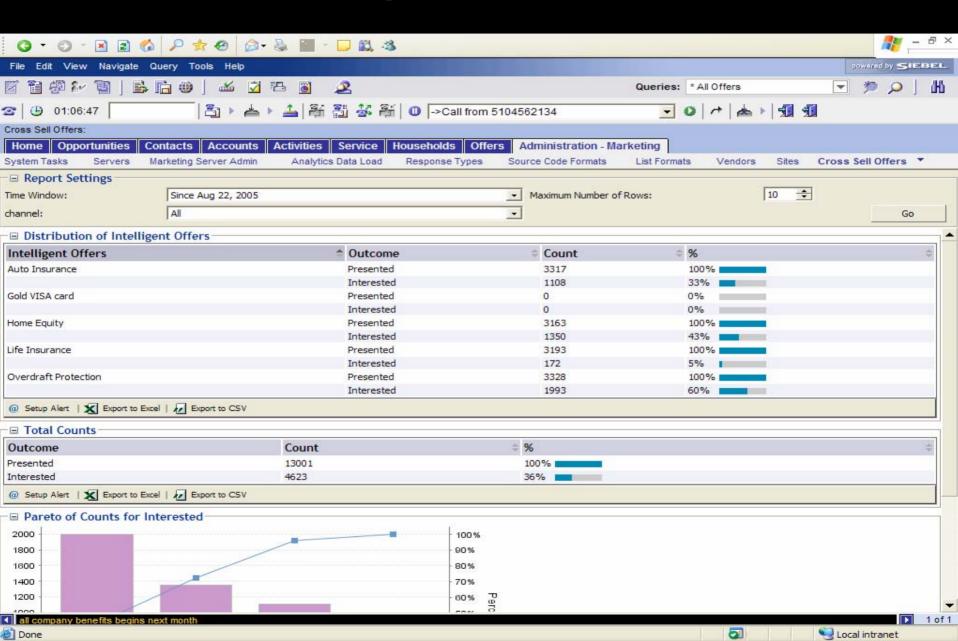
More in-depth reporting on key drivers



Trends in customer response behavior



Overall offer response statistics



Call Scenario B: Real-time Retention

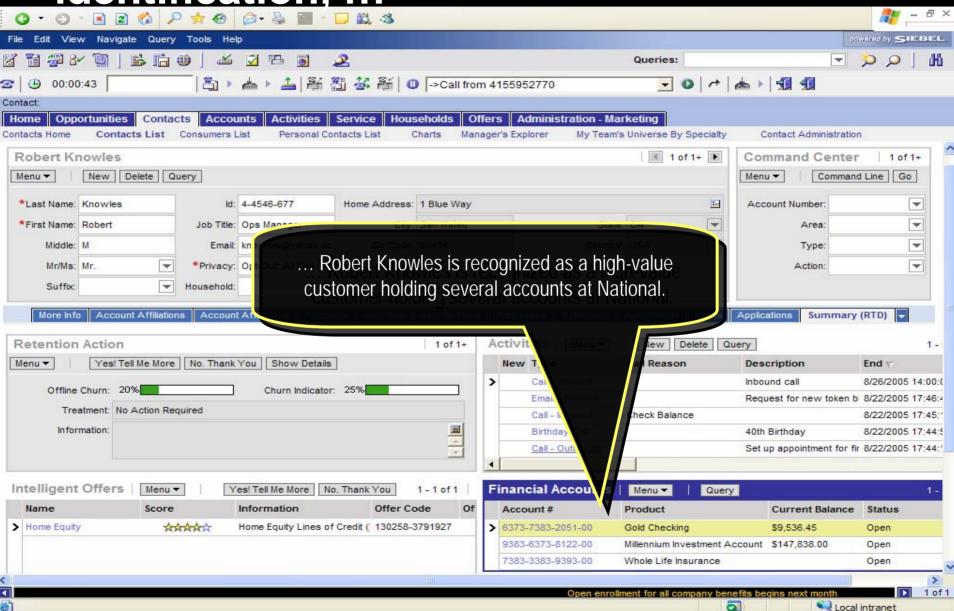
Profile of Caller (Robert Knowles):

- Male, 38 years old, married, homeowner
- Holds several accounts at National Bank
- High-value customer
- Considers closing all accounts (unknown to National Bank)
- Calls to inquire about checking account fees

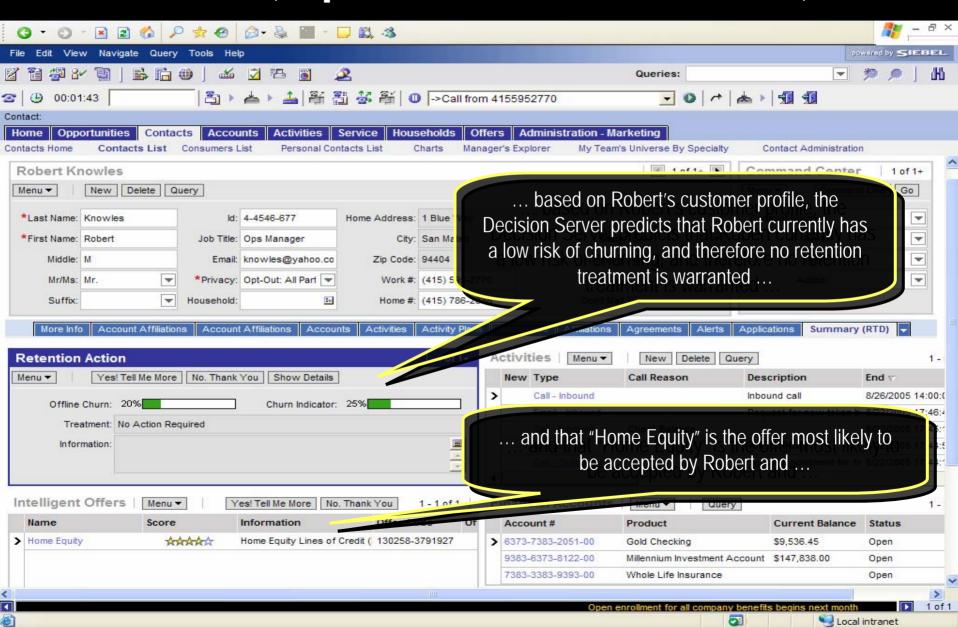
Objectives of National Bank:

Retain customer relationship through real-time retention treatment

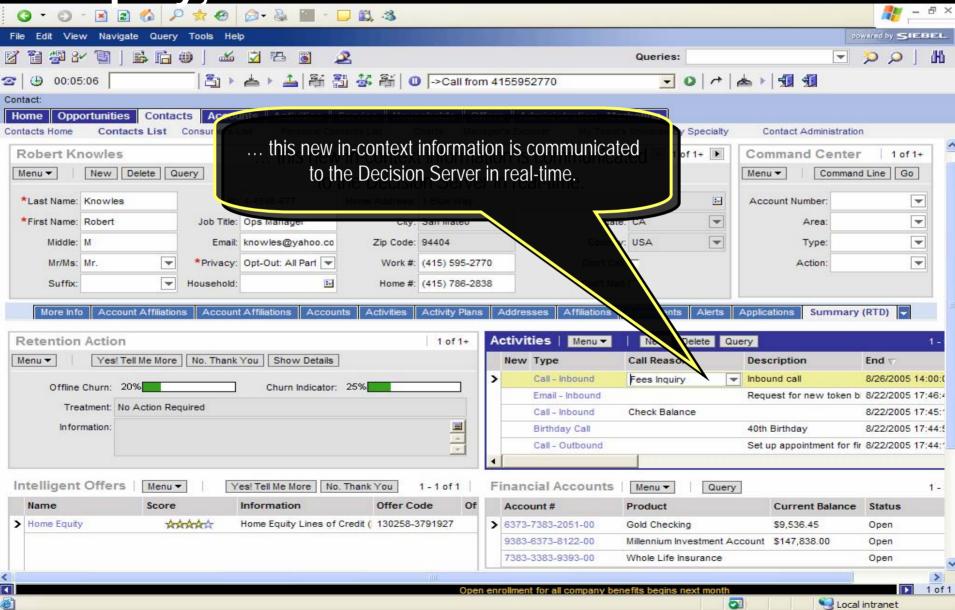
Call Scenario B: Upon caller identification, ...



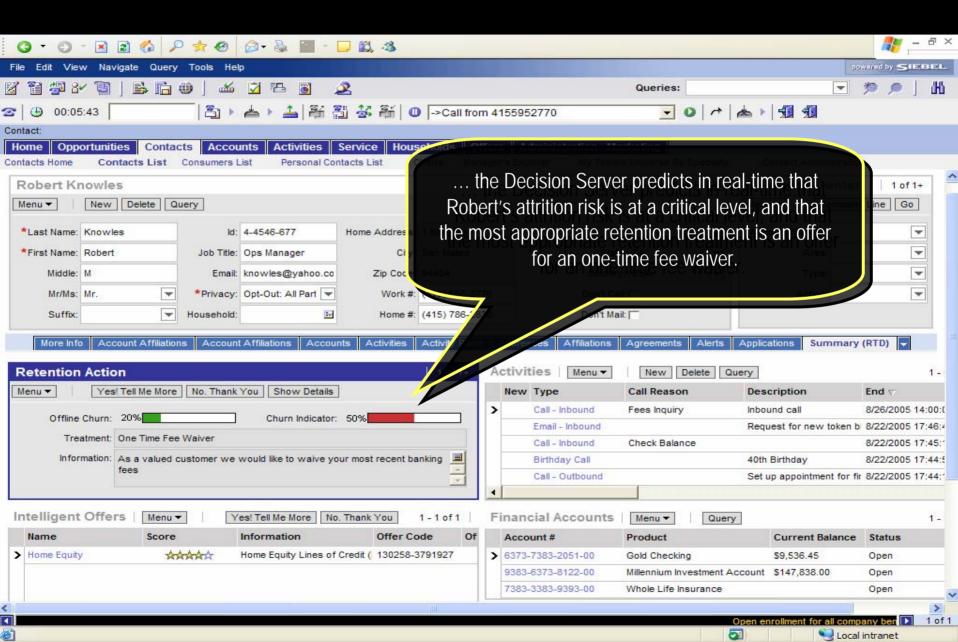
In addition, upon caller identification, ...



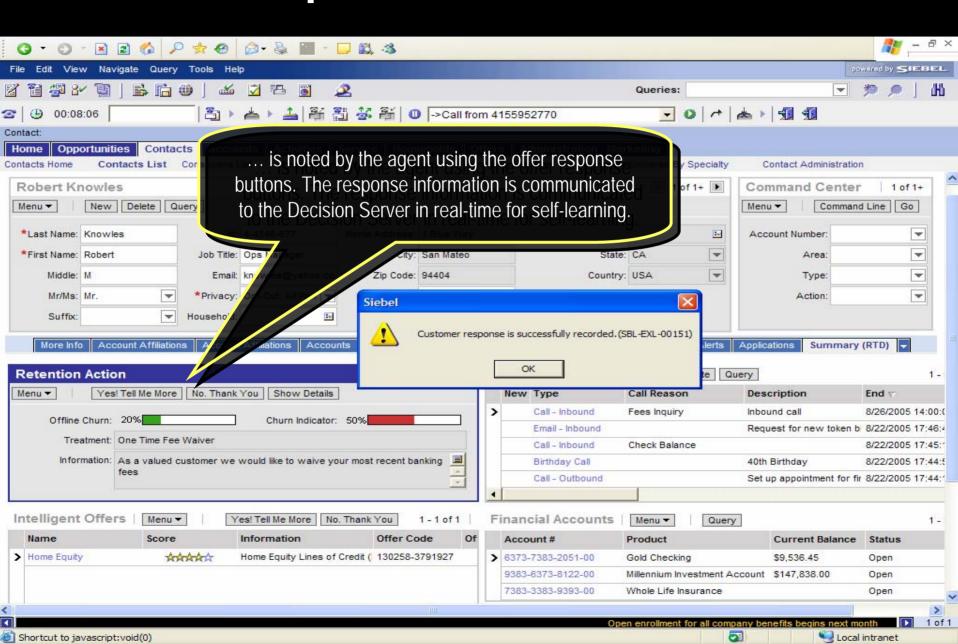
Upon noting Robert's call reason, fees inquiry, ...

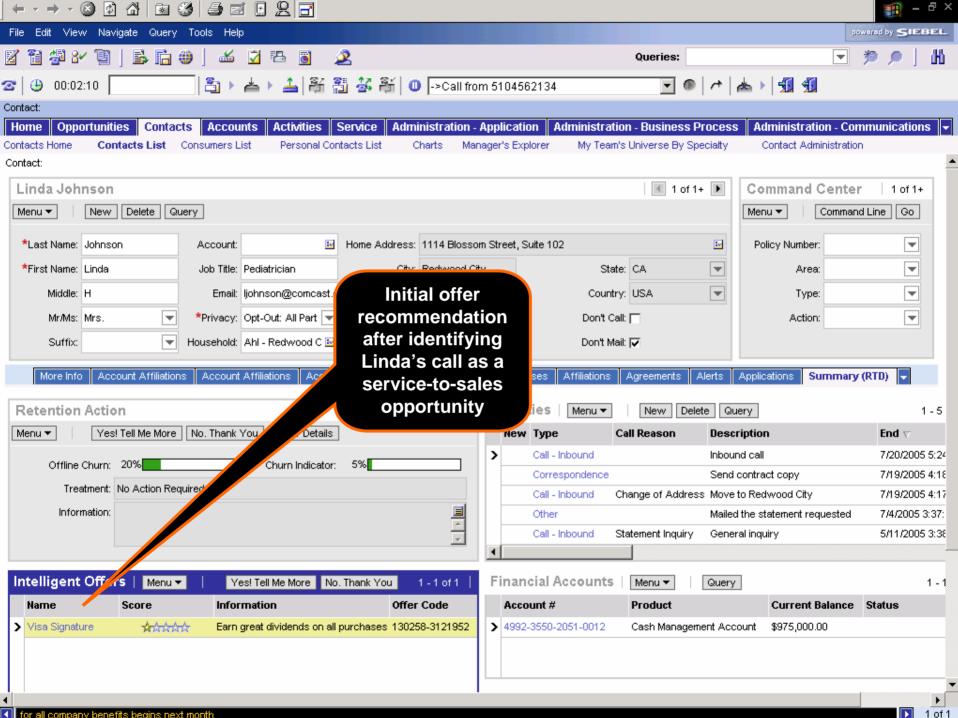


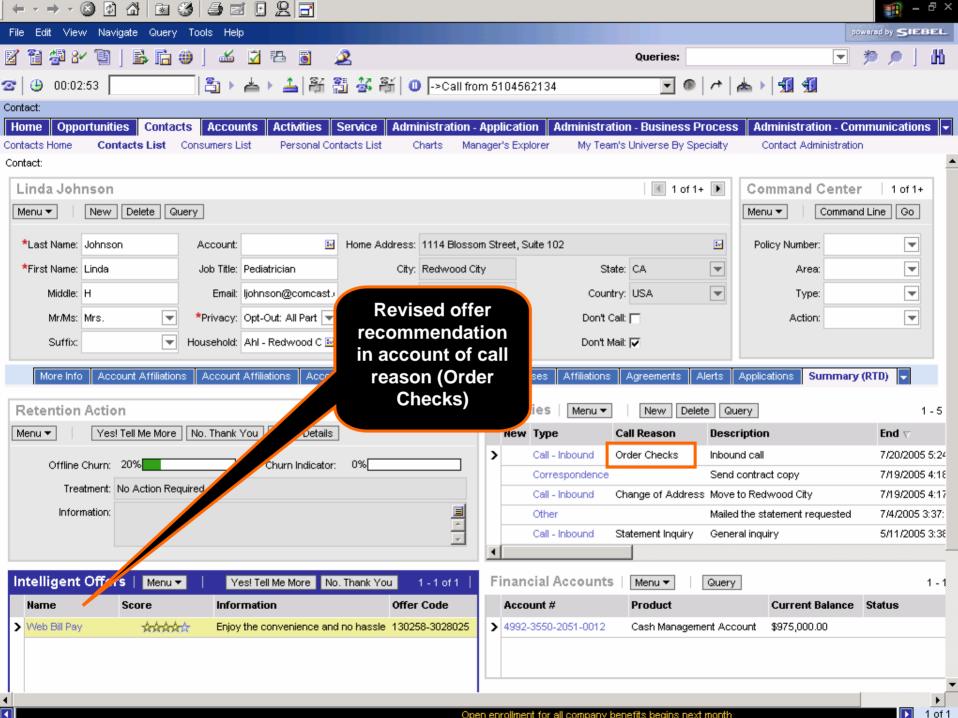
Based on the new in-context information, ...

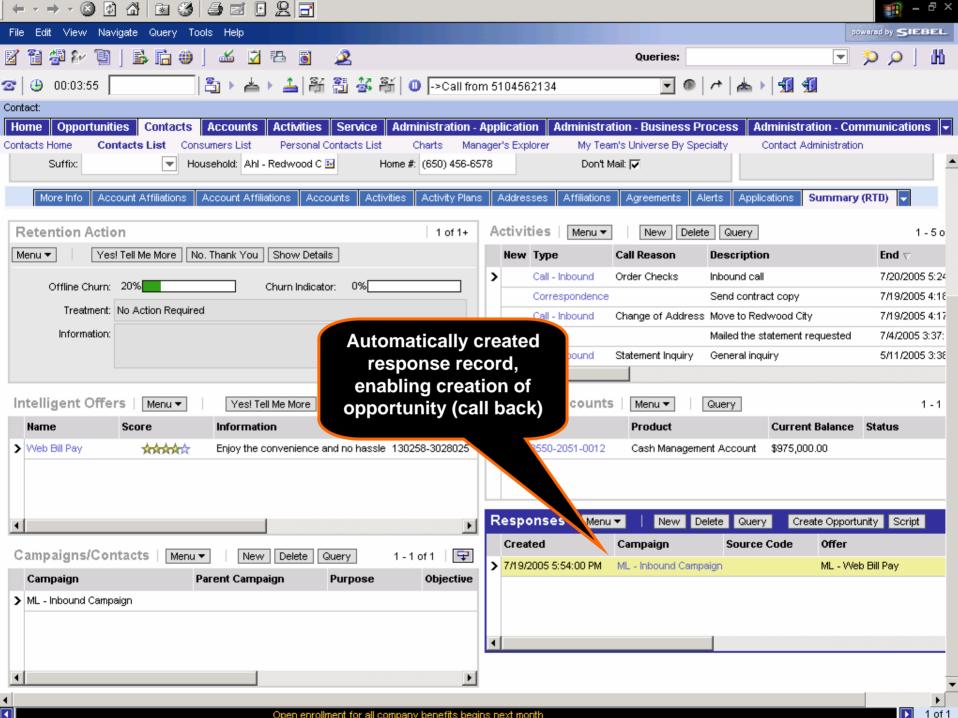


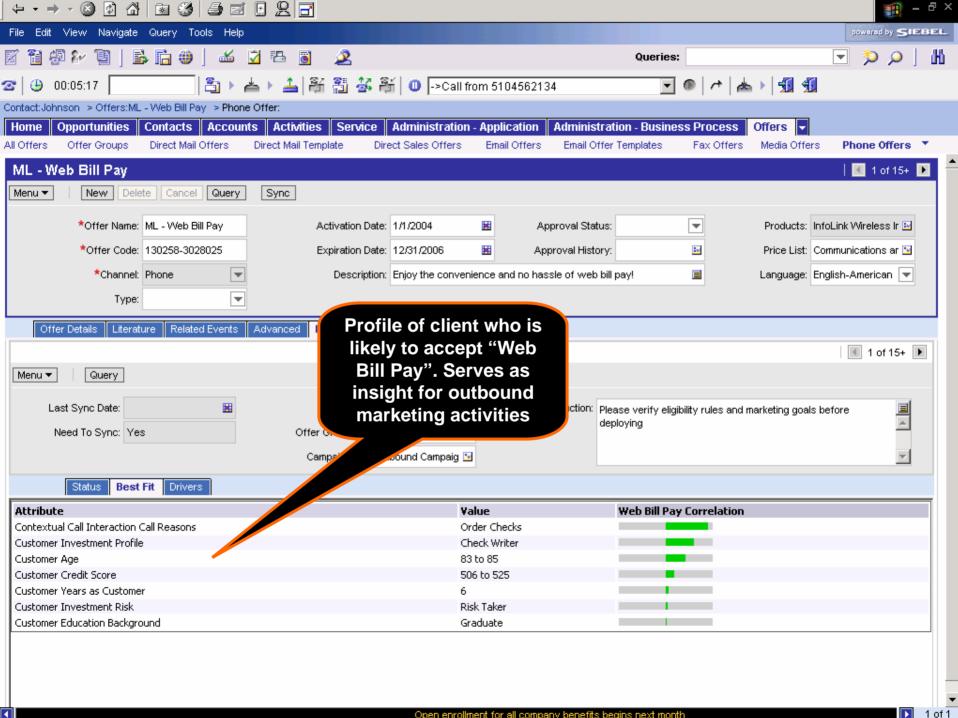
Robert's response to the extended offer ...

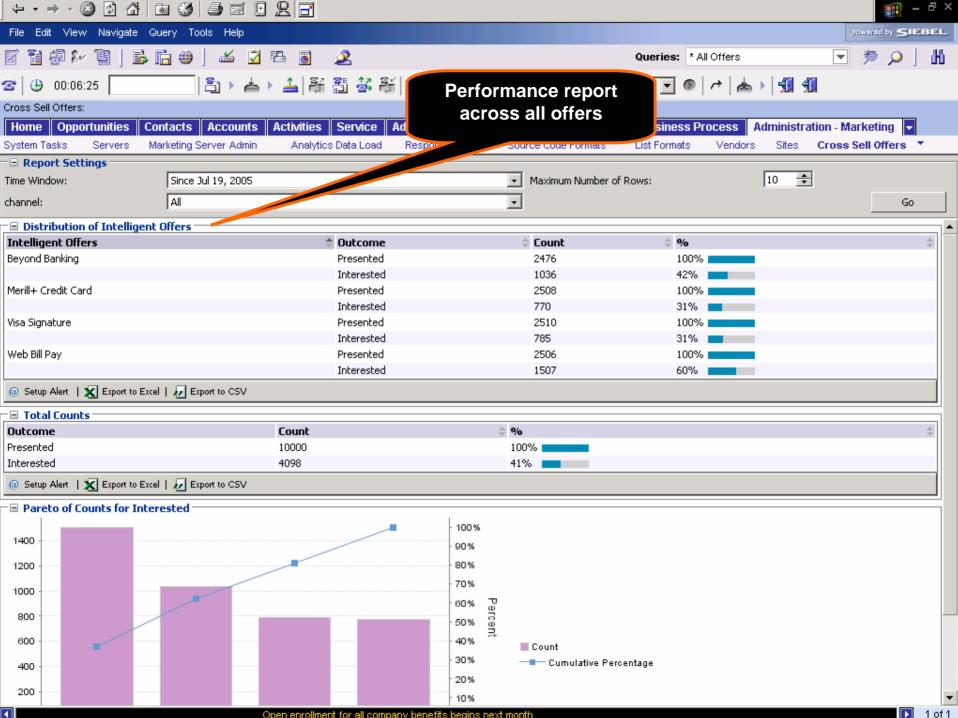


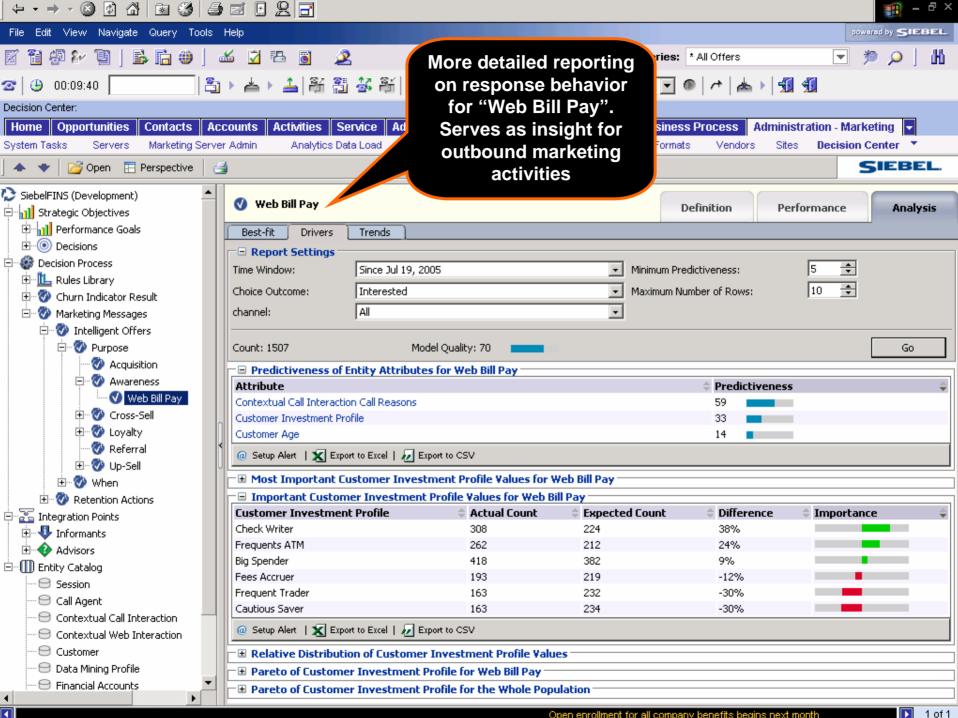


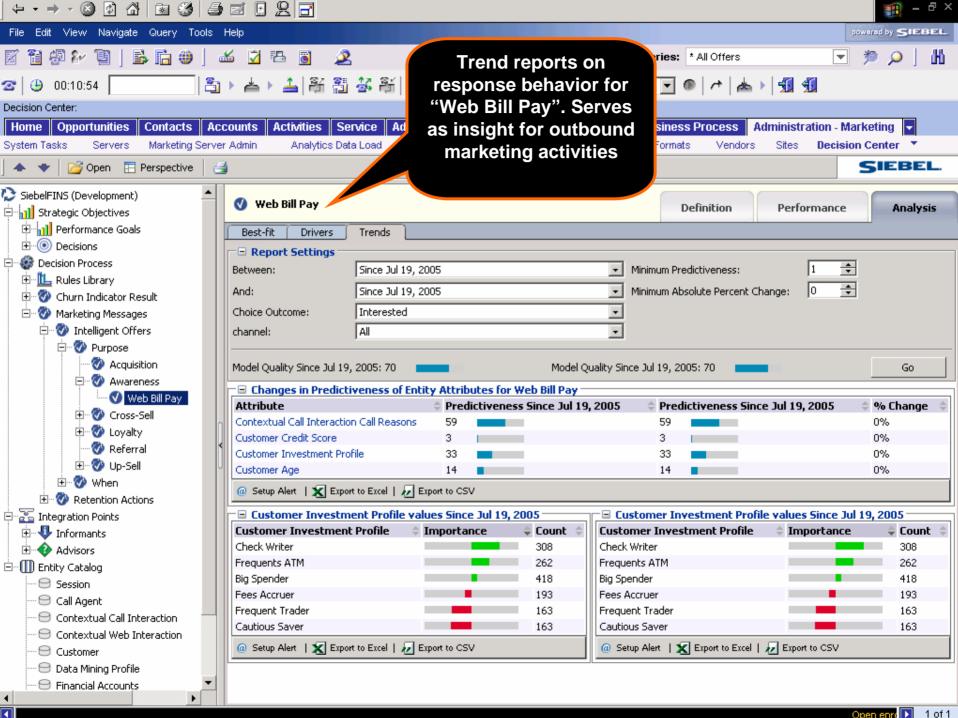


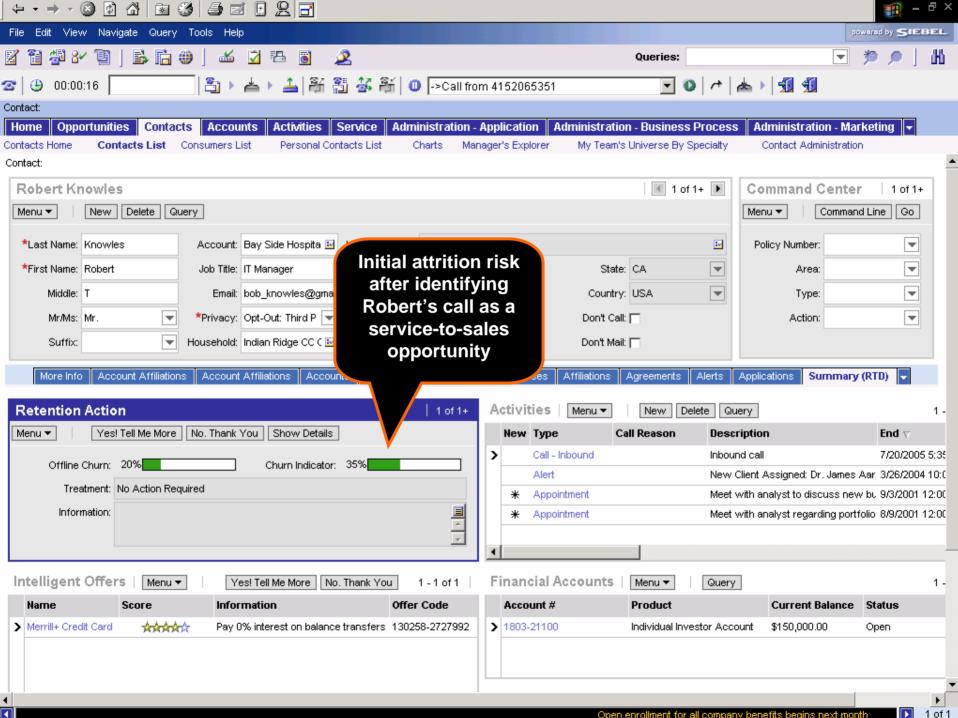


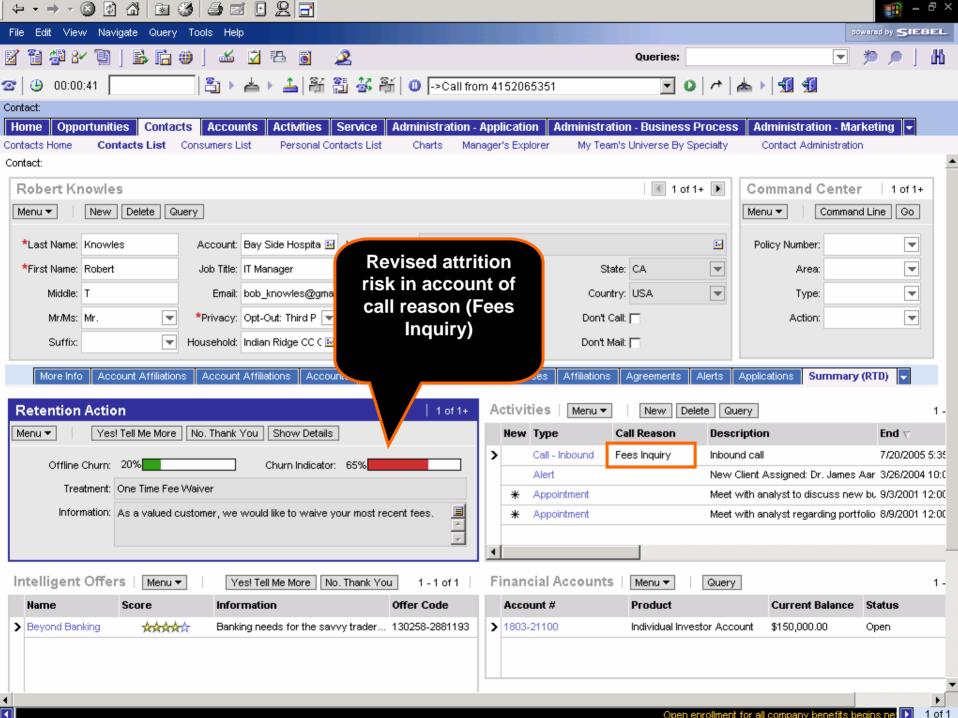


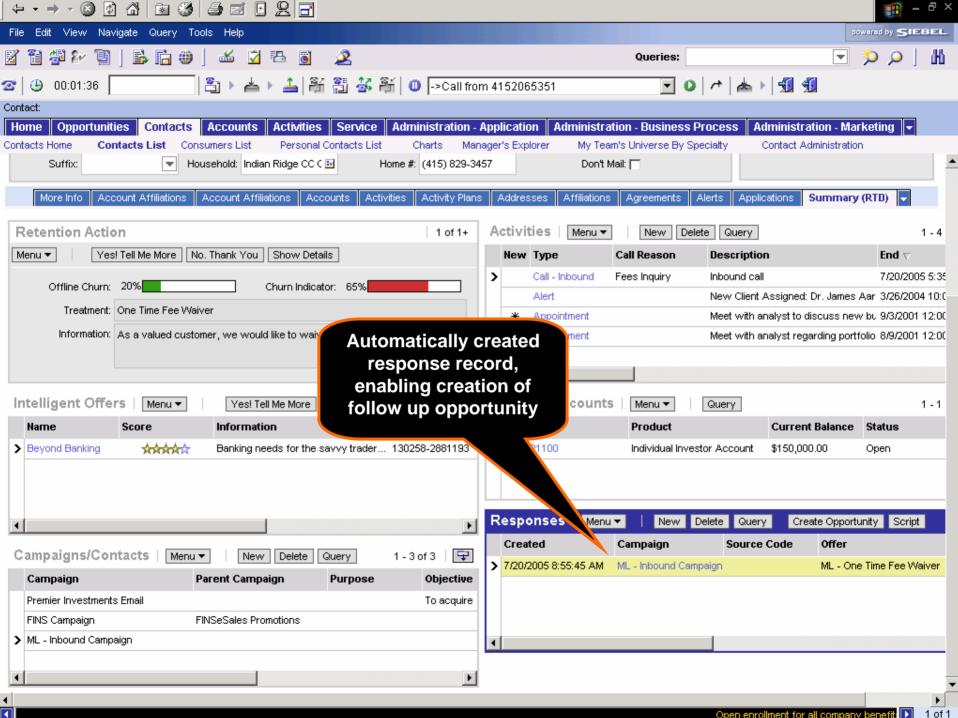












Real-Time Decision Framework





Increase Customer Revenue

Reduce Churn



Business Rules

High Value Customer = Annual Revenue > \$100 AND

Customer Tenure > 36 Months



Cross-Sell Decision			
Alternatives	Likely Success	Impact On Retention	Impact On Revenue
Gold Card	7%	+	+
Miles Card	17%	++	-
Platinum Card	3%	+	++

Optimal
Cross-Sell
Recommendation





Installing Oracle Data Mining

Architecture for Oracle Data Mining

- Installing Oracle Data Mining
- How to optimize the DB configuration?
- How to understand the schemas necessary to support Data Mining?

Installation Test

```
C:\WINDOWS\system32\cmd.exe - sqlplus wlan/wlan

C:\Documents and Settings\shyam>sqlplus wlan/wlan

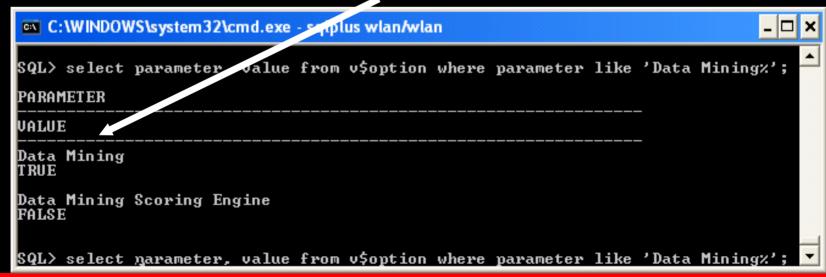
SQL*Plus: Release 10.2.0.1.0 - Production on Tue May 23 23:43:24 2006

Copyright (c) 1982, 2005, Oracle. All rights reserved.

Connected to:
Oracle Database 10g Enterprise Edition Release 10.2.0.1.0 - Production
With the Partitioning, Oracle Label Security, OLAP and Data Mining options

SQL>
```

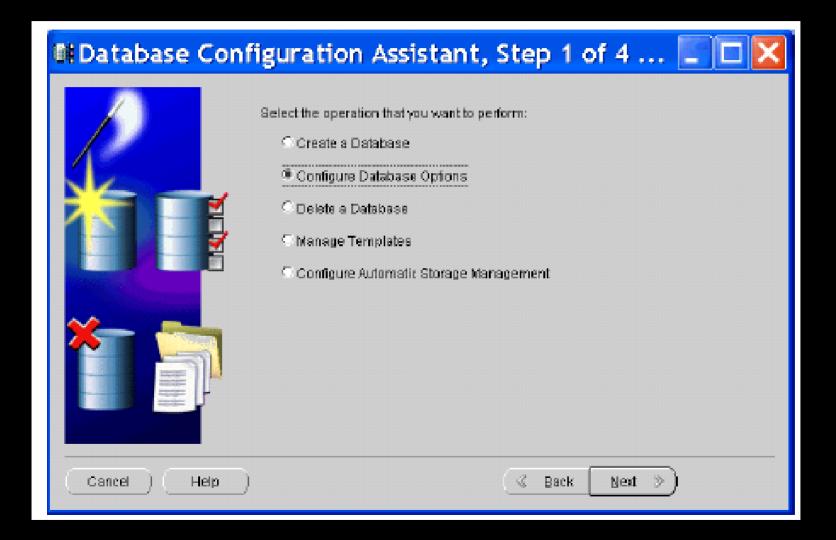
Check to see if the product exists.



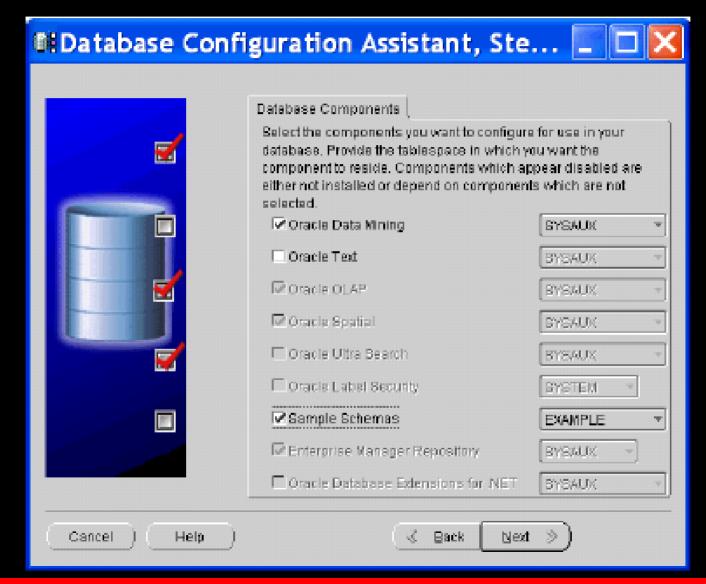
Install Option

- Default install list for Oracle 9i and 10g RDBMS include ODM
- It's a separate license, free for development, fee for production
- If ODM not installed, can be added by custom install using Oracle Universal Installer (OUI)
- Companion disk has samples programs for ODM
- ODMr the GUI
- ODM website

Add ODM



ODM...



Installing ODM GUI

- The GUI is called ODMr
- Uses JDBC connection to the database
- Can be used for most data mining tasks such as importing data from flat files, running models, scoring, visually viewing the results etc.





How to optimize the DB configuration

Sizing the DB

- Depends on use of Data warehouse or Data Mart as the ODM schema
- Size of the data to be used for mining
- Batch mode v/s real time response (analogyoptimize DB for total query run time v/s first response)

Config.ora Parameters

- If DB is used for other purposes, parallel DML can be disabled for the Data Mining sessions using the PL/SQL API's
 - alter session disable parallel query;
 - alter session disable parallel dml;
 - alter session disable parallel ddl;
- Some of the data mining sessions by-pass the underlying table's parallel settings
- Runs on RAC and Grid seamlessly





Schema for Data Mining

Creating mining users

- CREATE TABLESPACE "ODMPERM" DATAFILE
 'C:\ORACLE\PRODUCT\10.2.0\ORADATA\ORCL
 \odm1.dbf\' SIZE 20M REUSE AUTOEXTEND
 ON NEXT 20M;
- The next SQL command creates a new temporary tablespace.

```
CREATE TEMPORARY TABLESPACE "ODMTEMP"
TEMPFILE
'C:\ORACLE\PRODUCT\10.2.0\ORADATA\ORCL
\odmtemp.tmp' SIZE 20M REUSE
AUTOEXTEND ON NEXT 20M;
```

DB Privileges to the DM user

- Access Rights: Data mining users require several CREATE privileges. For text mining, users must also have access to the Oracle Text package ctxsys.ctx_ddl. The following privileges are required.
- CREATE PROCEDURE
- CREATE SESSION
- CREATE TABLE
- CREATE SEQUENCE
- CREATE VIEW
- CREATE JOB
- CREATE TYPE
- CREATE SYNONYM
- EXECUTE ON ctxsys.ctx ddl

SQL for Create user

- CREATE USER dmuser1 IDENTIFIED BY change_now DEFAULT TABLESPACE odmperm TEMPORARY TABLESPACE odmtemp QUOTA UNLIMITED on odmperm;
- SQL for Grants
 - GRANT create procedure to DMUSER1;
 - GRANT create session to DMUSER1;
 - GRANT create table to DMUSER1;
 - GRANT create sequence to DMUSER1; ...
- Export / import data mining models
 - SQL> EXECUTE
 DBMS_DATA_MINING.EXPORT_MODEL('allmodels.dmp','D
 MTEST');



Contact information:
Shyam.Nath@Oracle.com
(954) 609 2402 cell

DW & BI Special Interest Group





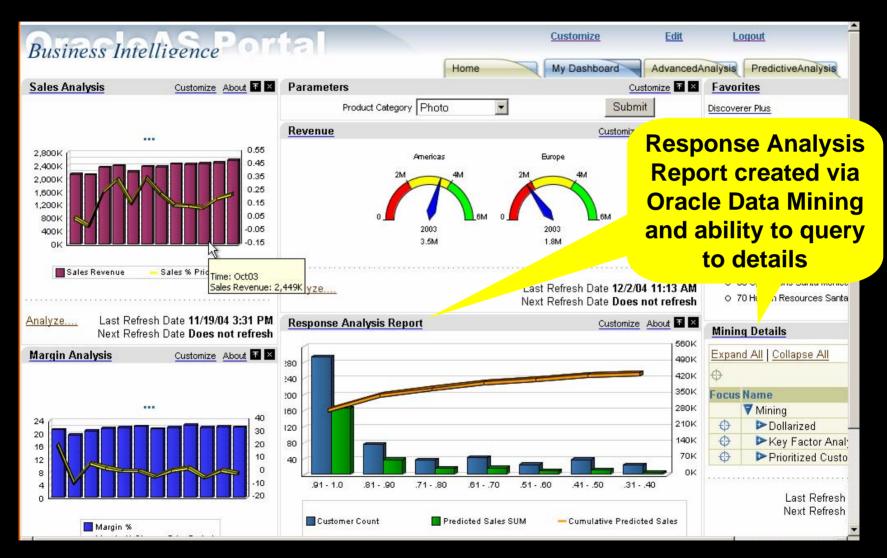


ENTERPRISE PLATFORM FOR

Business Intelligence

Oracle Business Intelligence

"Predictive Insights Reports"



Oracle Business Intelligence

"Predictive Insights Reports"

