ORACLE

Fraud and Anomaly Detection Using Oracle Advanced Analytic Option 12*c*

Charlie Berger
Sr. Director Product Management, Data Mining and Advanced Analytics
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine

Doctors, Nurses, Execs—Medicare Fraud

CNN International, By Terry Frieden, CNN Justice Producer, February 17, 2011 5:15 p.m. EST

- Federal authorities indicted and arrested more than 100 doctors, nurses and health care executives nationwide.
- Largest federal health care fraud takedown in our nation's history
  - The false billings to defraud Medicare totaled $225 million
- "From 2008 to 2010, every dollar the federal government spent under its health care fraud and abuse control programs averaged a return on investment (of) $6.80," Health and Human Services Secretary Kathleen Sebelius said.

American Society of Certified Fraud Examiners

20 Ways to Detect Fraud

1. Unusual Behavior

The perpetrator will often display unusual behavior, that when taken as a whole is a strong indicator of fraud. The fraudster may not ever take a vacation or call in sick in fear of being caught. He or she may not assign out work even when overloaded. Other symptoms may be changes in behavior such as increased drinking, smoking, defensiveness, and unusual irritability and suspiciousness.

2. Complaints

Frequently tips or complaints will be received which indicate that a fraudulent action is going on. Complaints have been known to be some of the best sources of fraud and should be taken seriously. Although at all times the motives of the complainant may be suspect, the allegations usually have merit that warrant further investigation.

3. Stale Items in Reconciliations

In bank reconciliations, deposits in transit, checks outstanding in the reconciliation could be indicative of theft. Missing deposits could mean the perpetrator is sitting on the funds, missing checks could indicate one made out to a bogus payee.

4. Excessive Voids

Voided sales slips could mean that the sale was rung up, the payment diverted to the use of the perpetrator, and the sales slip subsequently voided to cover the theft.

5. Missing Documents

Documents which are unable to be located can be a red flag for fraud. Although it is expected that some documents will be misplaced, the auditor should look for explanations as to why the documents are missing, and what steps were taken to locate the requested items. All too often, the auditors will select an alternate item or allow the auditee to select an alternate without determining whether or not a problem exists.

6. Excessive Credit Memos

Similar to excessive voids, this technique can be used to cover the theft of cash. A credit memo to a phony customer is written out, and the cash is taken to make total cash balance.

Pretty Easy? Huh?

http://www.fraudiscovery.com/detect.html

http://www.auditnet.org/testing_20_ways.htm

A Real Fraud Example

My credit card statement—Can you see the fraud?

	May 22	1:14 PM	FOOD	Monaco Café	$127.38
	May 22	7:32 PM	WINE	Wine Bistro	$28.00
	…	Gas Station?			
→	June 14	2:05 PM	MISC	Mobil Mart	$75.00
→	June 14	2:06 PM	MISC	Mobil Mart	$75.00
→	June 15	11:48 AM	MISC	Mobil Mart	$75.00
→	June 15	11:49 AM	MISC	Mobil Mart	$75.00
	May 28	6:31 PM	WINE	Acton Shop	$31.00
	May 29	8:39 PM	FOOD	Crossroads	$128.14
→	June 16	11:48 AM	MISC	Mobil Mart	$75.00
→	June 16	11:49 AM	MISC	Mobil Mart	$75.00

Total purchases exceeds time period average

Monaco?

Pairs of $75?

All same $75 amount?

Turkcell

Combating Communications Fraud

Objectives

- Prepaid card fraud—millions of dollars/year
- Extremely fast sifting through huge data volumes; with fraud, time is money

Solution

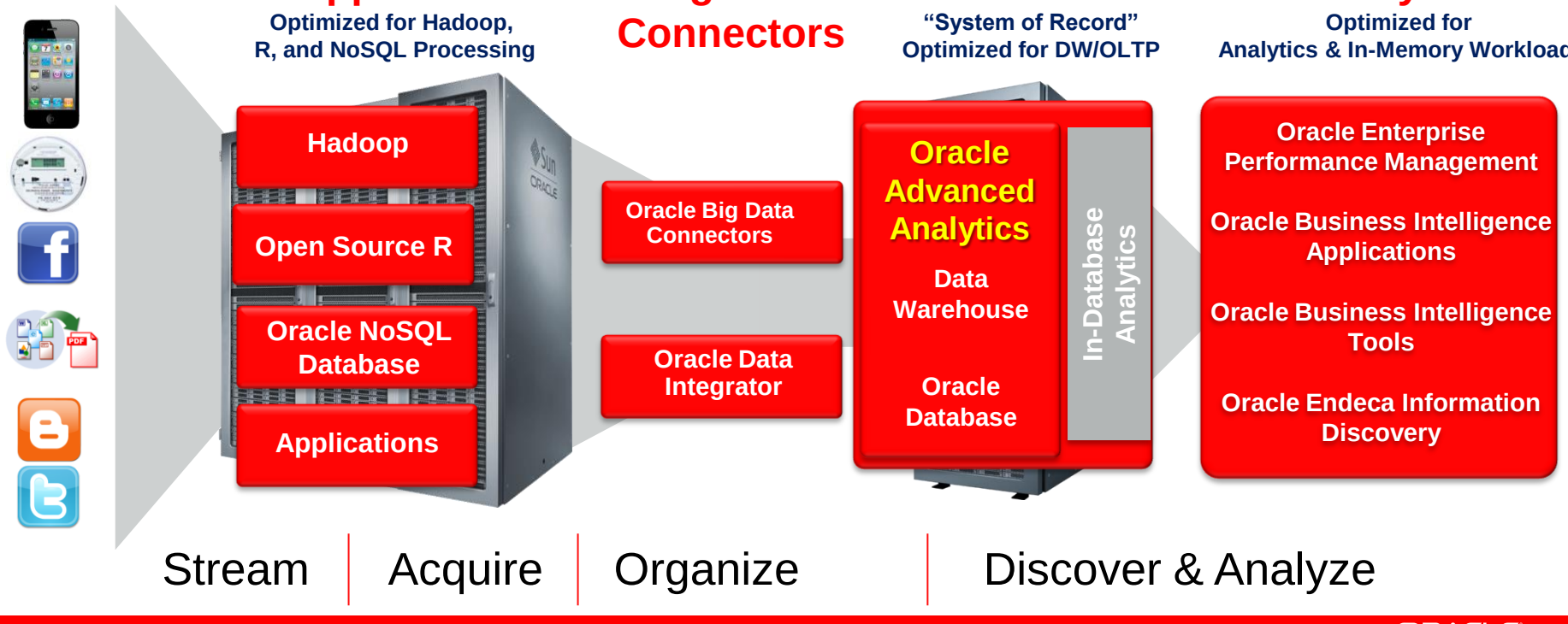
- Monitor 10 billion daily call-data records
- Leveraged SQL for the preparation—1 PB
- Due to the slow process of moving data, Turkcell IT builds and deploys models in-DB
- Oracle Advanced Analytics on Exadata for extreme speed. Analysts can detect fraud patterns almost immediately

- "Turkcell manages 100 terabytes of compressed data—or one petabyte of uncompressed raw data—on Oracle Exadata. With Oracle Data Mining, a component of the Oracle Advanced Analytics Option, we can analyze large volumes of customer data and call-data records easier and faster than with any other tool and rapidly detect and combat fraudulent phone use."
  – Hasan Tonguç Yılmaz, Manager, Turkcell İletişim Hizmetleri A.Ş.

Oracle Advanced Analytics
In-Database Fraud Models

Exadata

Oracle Big Data Platform



Oracle Advanced Analytics

Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- In-database data mining algorithms and open source R algorithms
- SQL, PL/SQL, R languages
- Scalable, parallel in-database execution
- Workflow GUI and IDEs
- Integrated component of Database
- Enables enterprise analytical applications

Oracle Advanced Analytics

Wide Range of In-Database Data Mining and Statistical Functions

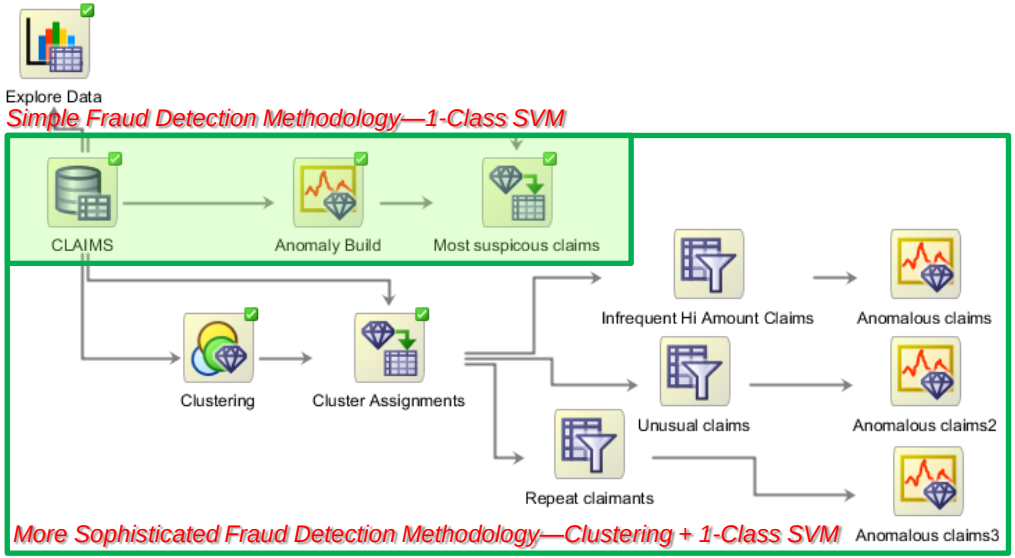
- Data Understanding & Visualization
  - Summary & Descriptive Statistics
  - Histograms, scatter plots, box plots, bar charts
  - R graphics: 3-D plots, link plots, special R graph types
  - Cross tabulations
  - Tests for Correlations (t-test, Pearson's, ANOVA)
  - Selected Base SAS equivalents
- Data Selection, Preparation and Transformations
  - Joins, Tables, Views, Data Selection, Data Filter, SQL time windows, Multiple schemas
  - Sampling techniques
  - Re-coding, Missing values
  - Aggregations
  - Spatial data
  - R to SQL transparency and push down
- Classification Models
  - Logistic Regression (GLM)
  - Naive Bayes
  - Decision Trees
  - Support Vector Machines (SVM)
  - Neural Networks (NNs)
- Regression Models
  - Multiple Regression (GLM)
  - Support Vector Machines
- Clustering
  - Hierarchical K-means
  - Orthogonal Partitioning
  - Expectation Maximization
- Anomaly Detection
  - Special case Support Vector Machine (1-Class SVM)
- Associations / Market Basket Analysis
  - A Priori algorithm
- Feature Selection and Reduction
  - Attribute Importance (Minimum Description Length)
  - Principal Components Analysis (PCA)
  - Non-negative Matrix Factorization
  - Singular Vector Decomposition
- Text Mining
  - Most OAA algorithms support unstructured data (i.e. customer comments, email, abstracts, etc.)
- Transactional Data
  - Most OAA algorithms support transactional data (i.e. purchase transactions, repeated measures over time)
- R packages—ability to run open source
  - Broad range of R CRAN packages can be run as part of database process via R to SQL transparency and/or via Embedded R mode

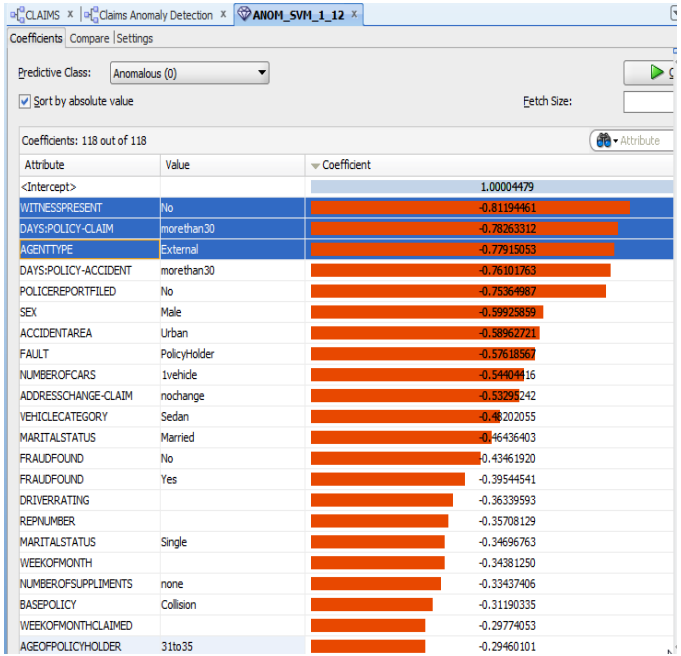
ORACLE

* included in every Oracle Database

Financial Sector/Accounting/Expenses

Anomaly Detection





Fraud Prediction Demo

Automated In-DB Analytical Methodolgies

```
drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values ('ALGO_NAME','ALGO_SUPPORT_VECTOR_MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;
```

```
begin
dbms_data_mining.create_model('CLAIMSMODEL', 'CLASSIFICATION',
  'CLAIMS', 'POLICYNUMBER', null, 'CLAIMS_SET');
end;
/
```

```
-- Top 5 most suspicious fraud policy holder claims
select * from
(select POLICYNUMBER, round(prob_fraud*100,2) percent_fraud,
     rank() over (order by prob_fraud desc) rnk from
(select POLICYNUMBER, prediction_probability(CLAIMSMODEL, '0' using *) prob_fraud
from CLAIMS
where PASTNUMBEROFCLAIMS in ('2to4', 'morethan4')))
where rnk <= 5
order by percent_fraud desc;
```

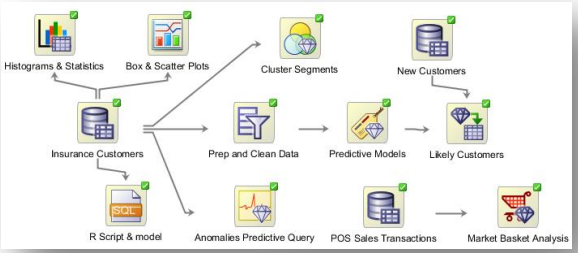
POLICYNUMBER	PERCENT_FRAUD	RNK
6532	64.78	1
2749	64.17	2
3440	63.22	3
654	63.1	4
12650	62.36	5

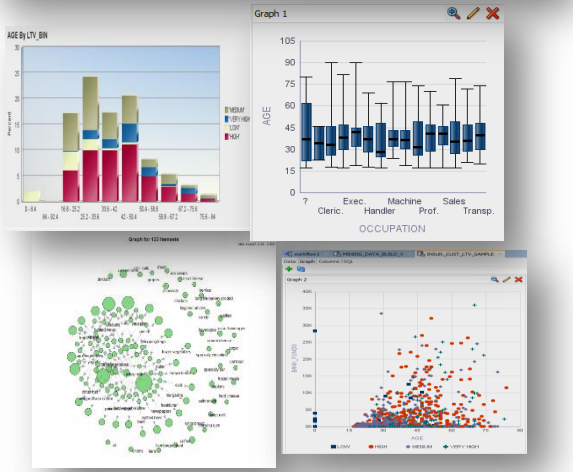
Automated Monthly “Application”! *Just add:*

```
Create
View CLAIMS2_30
As
Select * from CLAIMS2
Where mydate > SYSDATE – 30
```

Why Oracle Advanced Analytics?

Differentiating Features





- ✓ Fastest Way to Deliver Enterprise Predictive Analytics Applications
  - Integrated with OBIEE and any application that uses SQL queries
- ✓Performance and Scalability
  - Leverages power and scalability of Oracle Database.
- ✓Lowest Total Costs of Ownership
  - No need for separate analytical servers

A Real Fraud Example

My credit card statement—Can you see the fraud?

	May 22	1:14 PM	FOOD	Monaco Café	$127.38
	May 22	7:32 PM	WINE	Wine Bistro	$28.00
	…	Gas Station?			
→	June 14	2:05 PM	MISC	Mobil Mart	$75.00
→	June 14	2:06 PM	MISC	Mobil Mart	$75.00
→	June 15	11:48 AM	MISC	Mobil Mart	$75.00
→	June 15	11:49 AM	MISC	Mobil Mart	$75.00
	May 28	6:31 PM	WINE	Acton Shop	$31.00
	May 29	8:39 PM	FOOD	Crossroads	$128.14
→	June 16	11:48 AM	MISC	Mobil Mart	$75.00
→	June 16	11:49 AM	MISC	Mobil Mart	$75.00

Total purchases exceeds time period average

Monaco?

Pairs of $75?

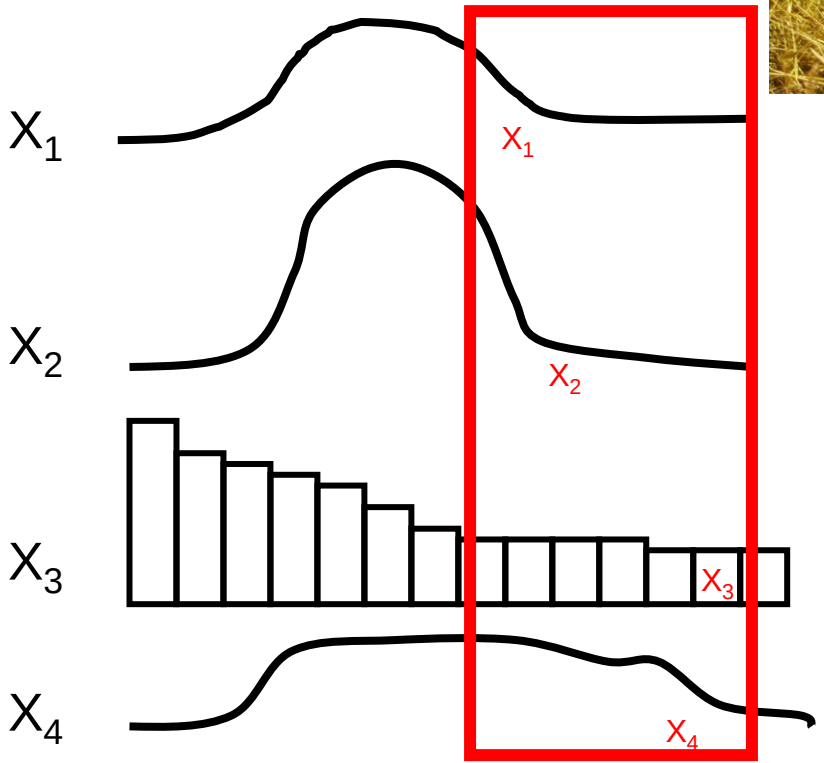
All same $75 amount?

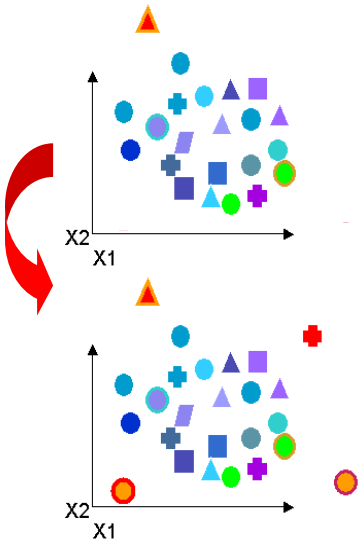
Multiple Approaches To Detect Potential Fraud

1. Anomaly Detection (1-Class SVM)
   - Add feedback loop to purify the input training data over time and improve model performance
2. Classification
   - IF you have a lot of examples (25% or more) of fraud on which to train/learn
3. Clustering
   - Find records that don't high very high probability to fit any particular cluster and/or lie in the outlier/edges of the clusters
4. Hybrid of #3 and then #1
   - Pre-cluster the records to create "similar" segments and then apply anomaly detection models for each cluster
5. Panel of Experts
   - i.e. 3 out of 5 models predict possibly anomalous above 40% or any 1 out of N models considers this record unusual

Challenge: Finding Anomalies

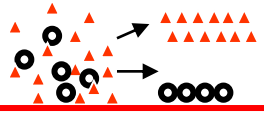
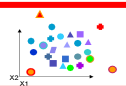
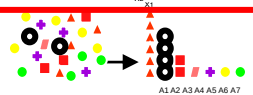
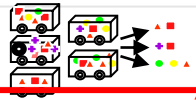
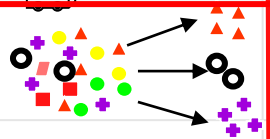
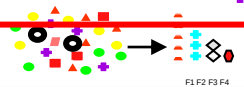
- Considering multiple attributes
- Taken alone, may seem “normal”
- Taken collectively, a record may appear to be anomalous
- Look for what is “*different*”





Oracle Advanced Analytics

SQL Data Mining Algorithms

Problem	Algorithms	Applicability
Classification 	• Logistic Regression (GLM) • Decision Trees • Naïve Bayes • Support Vector Machines	• Classical statistical technique • Popular for “rules” & transparency • Fast, simple, performant • New, versatile and performant
Regression 	• Multiple Regression (GLM) • Support Vector Machines	• Classical statistical technique • New, versatile and performant
Anomaly Detection 	• 1-Class Support Vector Machine	• Anomaly detection & fraud where lack examples of the target field
Attribute Importance 	• Minimum Description Length (MDL)	• Attribute reduction • Reduce data noise
Association Rules 	• Apriori	• Market basket analysis • Link analysis
Clustering 	• Hierarchical K-Means • Hierarchical O-Cluster • Expectation Maximization (EM)	• Customer segmentation • Find similar records, transactions or clusters
Feature Extraction 	• Principal Components Analysis (PCA) • Nonnegative Matrix Factorization • Singular Value Decomposition (SVD)	• Feature reduction e.g. many inputs, text problems, etc.

Oracle Advanced Analytics

Wide Range of In-Database Data Mining and Statistical Functions

- Data Understanding & Visualization
  - Summary & Descriptive Statistics
  - Histograms, scatter plots, box plots, bar charts
  - R graphics: 3-D plots, link plots, special R graph types
  - Cross tabulations
  - Tests for Correlations (t-test, Pearson's, ANOVA)
  - Selected Base SAS equivalents
- Data Selection, Preparation and Transformations
  - Joins, Tables, Views, Data Selection, Data Filter, SQL time windows, Multiple schemas
  - Sampling techniques
  - Re-coding, Missing values
  - Aggregations
  - Spatial data
  - R to SQL transparency and push down
- Classification Models
  - Logistic Regression (GLM)
  - Naive Bayes
  - Decision Trees
  - Support Vector Machines (SVM)
  - Neural Networks (NNs)
- Regression Models
  - Multiple Regression (GLM)
  - Support Vector Machines
- Clustering
  - Hierarchical K-means
  - Orthogonal Partitioning
  - Expectation Maximization
- Anomaly Detection
  - Special case Support Vector Machine (1-Class SVM)
- Associations / Market Basket Analysis
  - A Priori algorithm
- Feature Selection and Reduction
  - Attribute Importance (Minimum Description Length)
  - Principal Components Analysis (PCA)
  - Non-negative Matrix Factorization
  - Singular Vector Decomposition
- Text Mining
  - Most OAA algorithms support unstructured data (i.e. customer comments, email, abstracts, etc.)
- Transactional Data
  - Most OAA algorithms support transactional data (i.e. purchase transactions, repeated measures over time)
- R packages—ability to run open source
  - Broad range of R CRAN packages can be run as part of database process via R to SQL transparency and/or via Embedded R mode

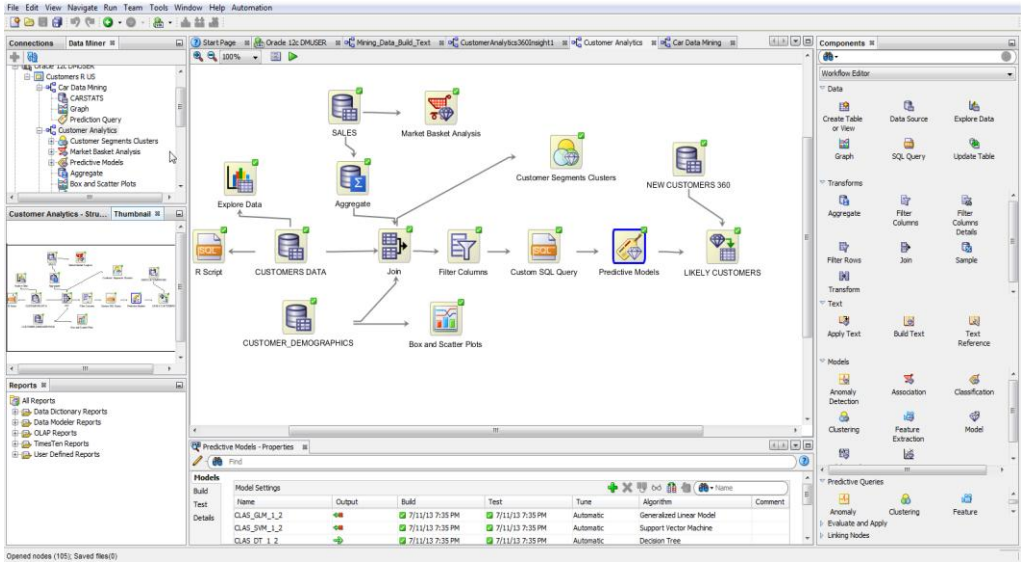
ORACLE

* included in every Oracle Database

Oracle Data Miner GUI

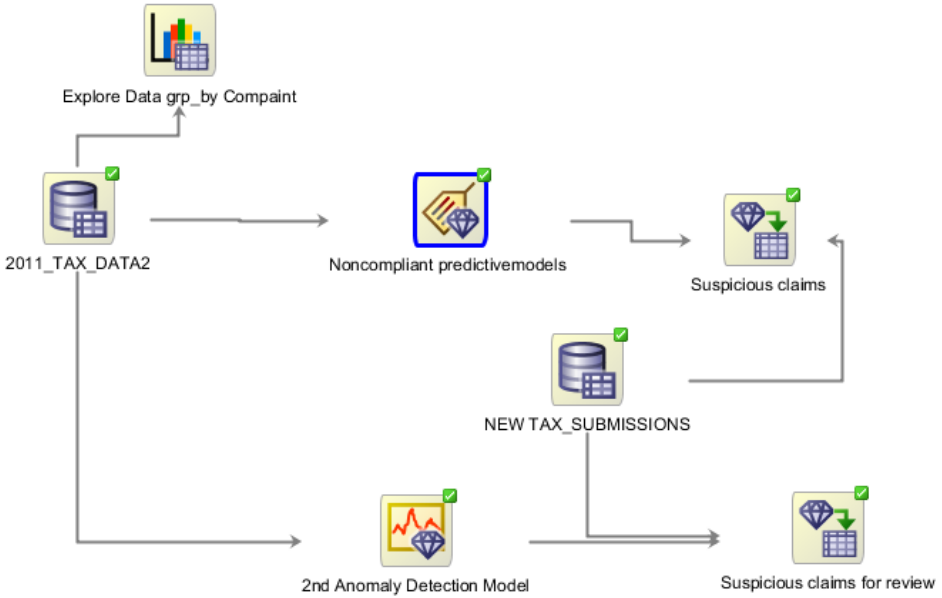
SQL Developer 4.0 Extension—*Free OTN Download*

- Easy to Use
  - Oracle Data Miner GUI for data analysts
  - “Work flow” paradigm
- Powerful
  - Multiple algorithms & data transformations
  - Runs 100% in-DB
  - Build, evaluate and apply models
- Automate and Deploy
  - Generate SQL scripts for deployment
  - Share analytical workflows



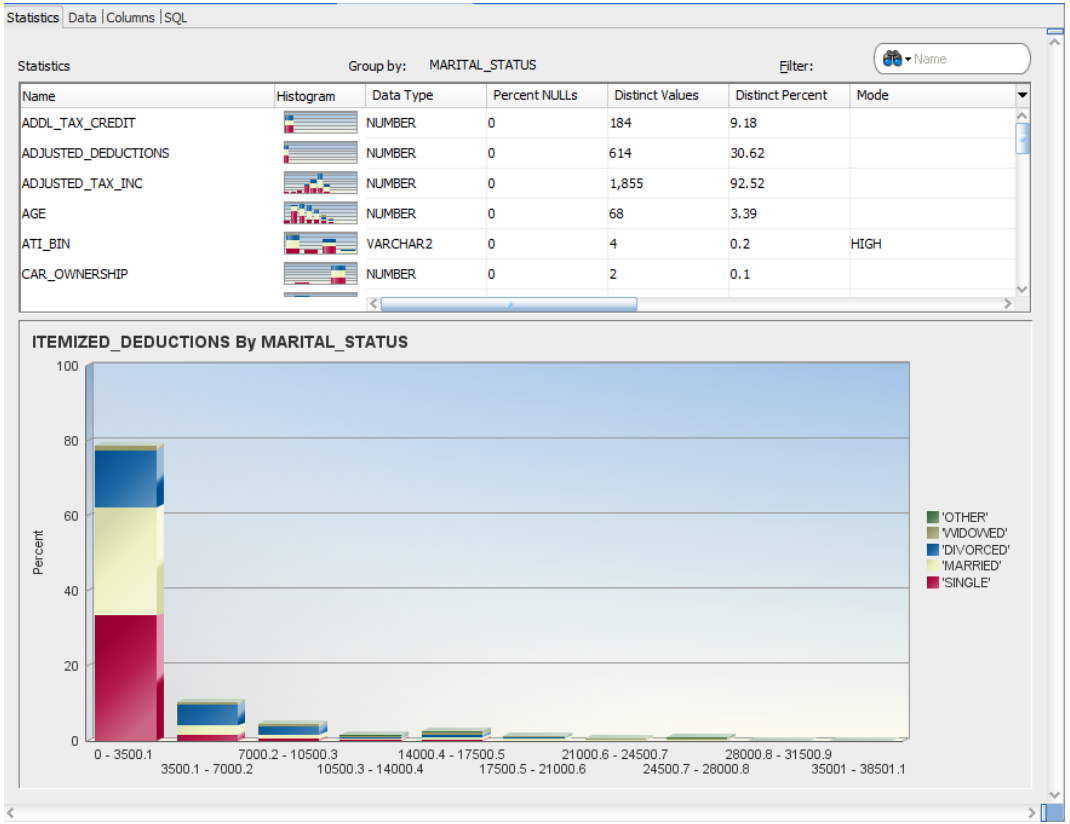
Tax Noncomplaince Audit Selection

- Simple Oracle Data Mining predictive model
  - Uses Decision Tree for classification of Noncompliant tax submissions (yes/no) based on historical 2011 data

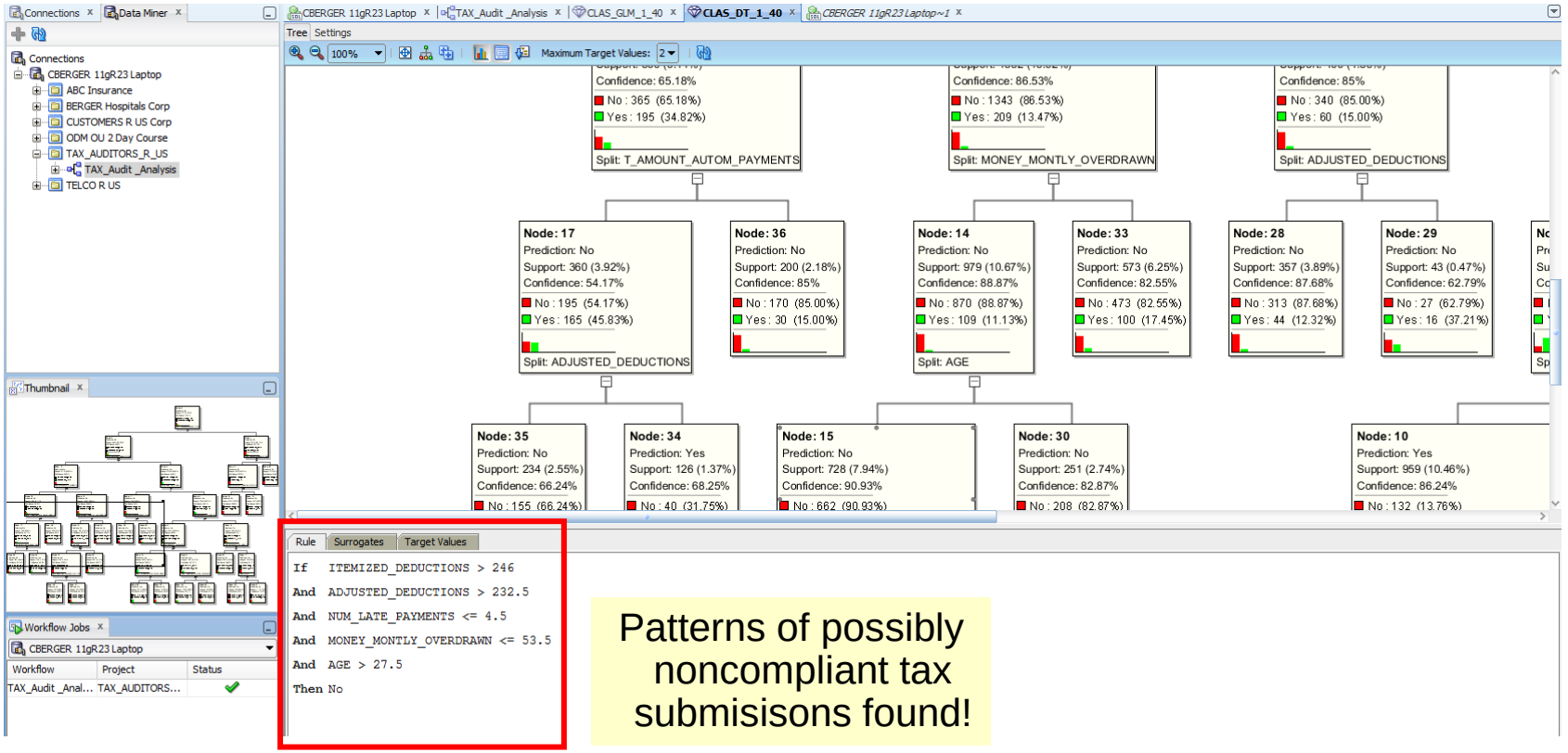


Tax Noncomplaince Audit Selection

- Tax data used for demo



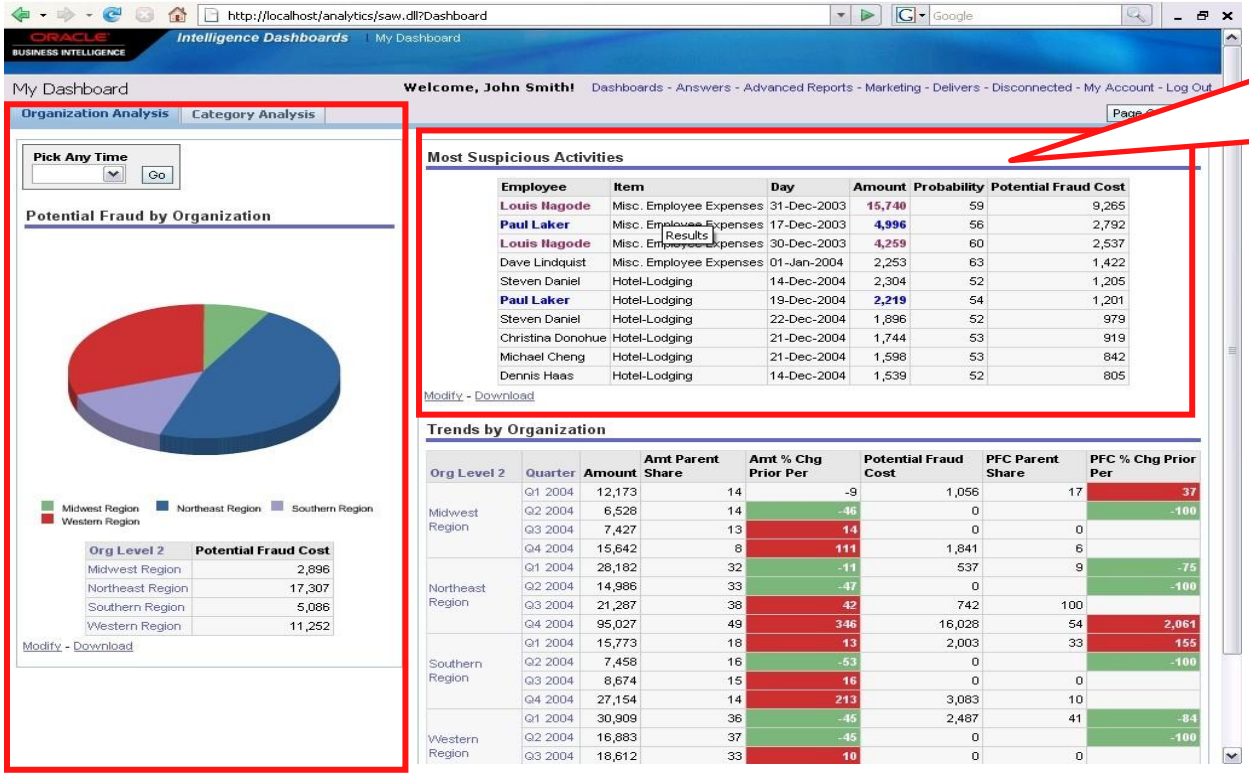
Tax Noncomplaince Audit Selection



Patterns of possibly noncompliant tax submisisons found!

Fraud and Non-Compliance Example

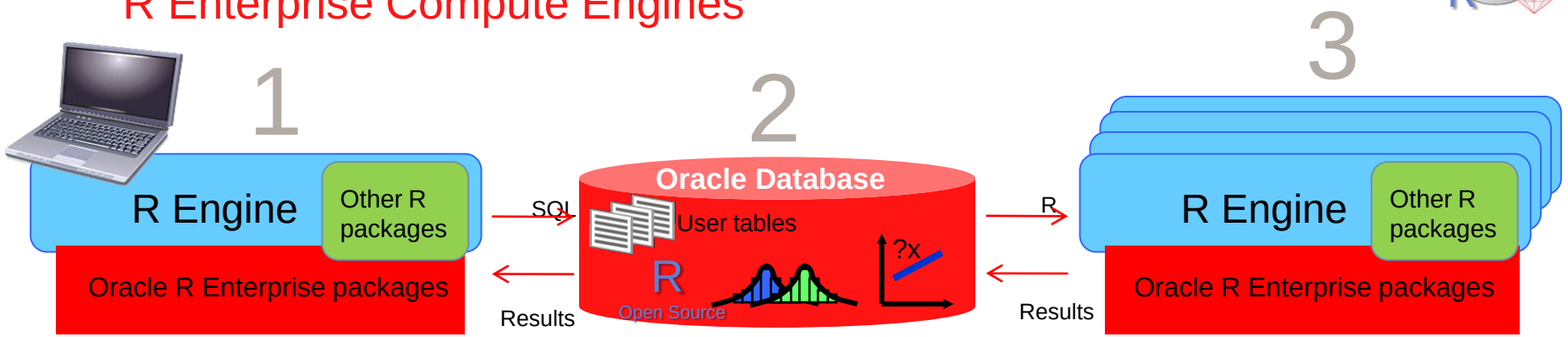
Identify & Drill-Thru Expenses by Probability of Non-Compliance



OAA data mining models provide likelihood of expense reporting fraud …and other important insights.

Oracle Advanced Analytics

R Enterprise Compute Engines



User R Engine on desktop

- R-SQL Transparency Framework intercepts R functions for scalable in-database execution
- Function intercept for data transforms, statistical functions and advanced analytics
- Interactive display of graphical results and flow control as in standard R
- Submit entire R scripts for execution by database

Database Compute Engine

- Scale to large datasets
- Access tables, views, and external tables, as well as data through DB LINKS
- Leverage database SQL parallelism
- Leverage new and existing in-database statistical and data mining capabilities

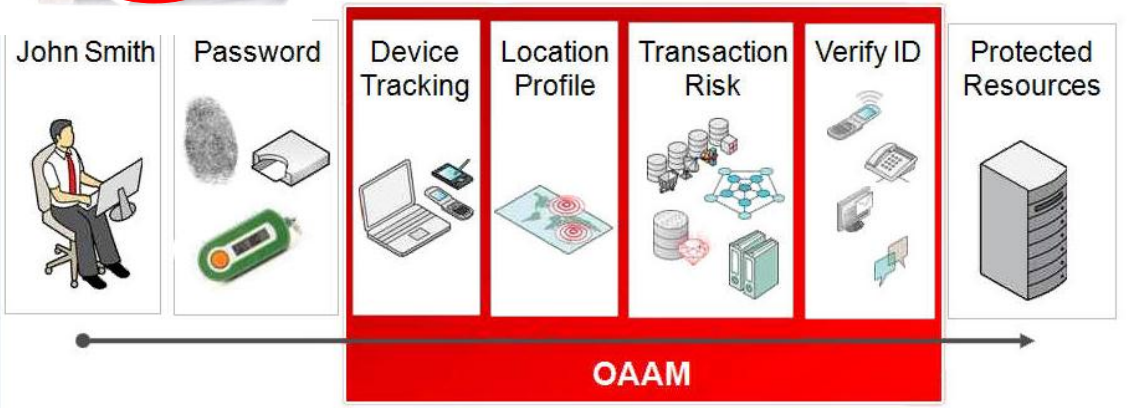
R Engine(s) spawned by Oracle DB

- Database can spawn multiple R engines for database-managed parallelism
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables “lights-out” execution of R scripts

Oracle Adaptive Access Manager

Trust…., But Verify

- Global ODM clustering model identifies typical behaviors/patterns/ profiles
- Each user is assigned several cluster nodes, that in total, capture 85% of their typical behavior/profile
- Real-time "*scoring*" of ODM model to bolster OAA's complex real-time security



Authentication is valid but is this really John Smith?

Is anything suspicious about John's transactions?

Can John answer a challenge if the risk is high?

Financial Sector/Accounting/Expenses

Oracle Spend Classification: Auto—Classify Spend into Purchasing Categories

- Text mining of expense items descriptions
- "Defragmentation" of likely misclassified expenses
  - "Flat panel monitor" ≠ "Meals"