Big Data Predictive Analytics in Oracle Database 12c

Oracle Advanced Analytics Database Option–Extending the Database to an Analytical Database

Charlie Berger, MS Eng, MBA
Sr. Director Product Management, Data Mining and Advanced Analytics
charlie.berger@oracle.com    www.twitter.com/CharlieDataMine
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Big Data is Big Business
Sources: The Economist, McKinsey & Company, Gartner, Facebook, IBM

- Every day, we create 2.5-quintillion bytes of data.
- 90 per cent of the data in the world today has been created in the past two years.
- Every minute, 100,000 tweets are sent globally.
- Google receives two-million search requests every minute.
- Five-billion mobile phones were in use in 2010.
- 30-billion pieces of content are shared on Facebook every month.
- By one estimate, there will be 5,200 gigabytes of data for every human on the planet by 2020.
- By 2015, 4.4-million IT jobs globally will be created to support big data, generating 1.9-million IT jobs in the United States alone.
- 70 per cent of data is created by individuals – but enterprises are responsible for storing and managing 80 per cent of it.
- Big data will drive $232 billion in spending through 2016.
- There is the potential for a 60 per cent increase in retailers’ operating margins with big data.

Analytics and Big Data
Are Changing the World

**ANALYTICS**

Unlock value in data to solve some of the world’s most pressing problems

**Transform Communities**
Using analytics and big data, U.N. officials will deliver energy to 1.3 billion people

**Save Lives**
Scientists protect consumers by pulling deadly medication off the market

**Increase Public Safety**
Police increase public safety by predicting crime “hot spots” and pre-deploying officers

**ORGANIZATIONS WHICH USE ANALYTICS GET**

$10.66 FOR EVERY $1 THEY SPEND ON ANALYTICS

**CHANGE YOUR BUSINESS**

Imagine what analytics can do for your business

3x Top performers are 3x more likely to use analytics than low performers

53% use analytics to drive strategy

50% use analytics to transform daily operations

http://www.pinterest.com/pin/267753140321592721/
Planning for Future
Growth of Data Exponentially Greater than Growth of Data Analysts!

Growth of Data vs. Growth of Data Analysts

Conclusion
– Data Analysis platforms need to be
  • Extremely Easy to Learn, yet..
  • Extremely Powerful and
  • Automated as much as possible!

Stored Data accumulating at 28% annual growth rate
Data Analysts in workforce growing at 5.7% growth rate

Data Analyst shortage

http://www.delphianalytics.net/more-data-than-analysts-the-real-big-data-problem/
Four Main Languages for Analytics, Data Mining, Data Science: R, SQL, SAS, Python

SQL is at Core

• BIG DATA’S BIG FLIP-FLOP
  – BY BILL FRANKS, Chief Analytics Officer for Teradata, MAR 13, 2014
  – “It wasn’t too long ago that many people espoused the decline, if not death, of the SQL language and relational database technology in general.”
  – “In case you hadn’t noticed, a huge flip-flop has occurred. Many of the same people and organizations that were recently dismissing the entire concept of relational environments and SQL are now racing to ... wait for it ... add SQL-style interfaces on top of non-relational platforms like Hadoop!”
Oracle Big Data Platform

**Oracle Big Data Appliance**
- Optimized for Hadoop, R, and NoSQL Processing
  - Hadoop
  - Oracle NoSQL Database
  - Open Source R
  - Applications

**Oracle Big Data Connectors**
- Oracle Big Data Connectors
  - Oracle Data Integrator

**Oracle Exadata**
- “System of Record” Optimized for DW/OLTP
  - Oracle Database
  - Data Warehouse
  - Oracle Advanced Analytics

**Oracle Exalytics**
- Optimized for Analytics & In-Memory Workloads
  - Enterprise Performance Management
  - Oracle BI Applications
  - Oracle BI EE
  - Endeca Information Discovery

**Stream** | **Acquire** | **Organize** | **Discover & Analyze**
Oracle Advanced Analytics + Exadata + Big Data SQL

OAA data mining model “scoring” pushed to Exadata storage tier and BDA

• With Oracle Advanced Analytics, SQL predicates and predictive models get pushed down for execution
  – For Exadata environments, get pushed to Exadata storage level for execution
  – For BDA environments, get pushed to BDA for execution

  – For example, find the US customers likely to churn:

    ```sql
    select cust_id
    from customers
    where region = 'US'
    and prediction_probability(churnmod, 'Y' using *) > 0.8;
    ```
Oracle Advanced Analytics Database Option
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 Database Evolution


• 7 Data Mining “Partners”

1998

• Oracle acquires Thinking Machine Corp’s dev. team + “Darwin” data mining software

1999

• Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API

2002

• Oracle Data Mining 10g & 10gR2 introduces SQL dm functions, 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI

2004

• New algorithms (EM, PCA, SVD)
• Predictive Queries
• ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
• SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
• OAA/ORE 1.3 + 1.4
• Integration with “R” and introduction/addition of Oracle R Enterprise
• Product renamed “Oracle Advanced Analytics (ODM + ORE)”
• New algorithms (EM, PCA, SVD)
• Predictive Queries
• SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
• OAA/ORE 1.3 + 1.4
• Integration with “R” and introduction/addition of Oracle R Enterprise
• Product renamed “Oracle Advanced Analytics (ODM + ORE)”

2005

• Oracle Data Mining 10g & 10gR2 introduces SQL dm functions, 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI

2008

• New algorithms (EM, PCA, SVD)
• Predictive Queries
• ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
• SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
• OAA/ORE 1.3 + 1.4
• Integration with “R” and introduction/addition of Oracle R Enterprise
• Product renamed “Oracle Advanced Analytics (ODM + ORE)”

2011

• New algorithms (EM, PCA, SVD)
• Predictive Queries
• ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
• SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
• OAA/ORE 1.3 + 1.4
• Integration with “R” and introduction/addition of Oracle R Enterprise
• Product renamed “Oracle Advanced Analytics (ODM + ORE)”

2014

• New algorithms (EM, PCA, SVD)
• Predictive Queries
• ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
• SQLDEV/Oracle Data Miner 4.0 SQL script generation and SQL Query node (R integration)
• OAA/ORE 1.3 + 1.4
• Integration with “R” and introduction/addition of Oracle R Enterprise
• Product renamed “Oracle Advanced Analytics (ODM + ORE)”

Oracle Advanced Analytics

Performance and Scalability with Low Total Cost of Ownership

Data remains in the Database
- Scalable, parallel Data Mining algorithms in SQL kernel
- Fast parallelized native SQL data mining functions, SQL data preparation and efficient execution of R open-source packages
- High-performance parallel scoring of SQL data mining functions and R open-source models

Fastest way to deliver enterprise-wide predictive analytics
- Integrated GUI for Predictive Analytics
- Database scoring engine

Lowest TCO
- Eliminate data duplication
- Eliminate separate analytical servers
- Leverage investment in Oracle IT
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.
More Data Variety—Better Predictive Models

• Increasing sources of relevant data can boost model accuracy

Model with “Big Data” and hundreds -- thousands of input variables including:
• Demographic data
• Purchase POS transactional data
• “Unstructured data”, text & comments
• Spatial location data
• Long term vs. recent historical behavior
• Web visits
• Sensor data
• etc.

![Graph showing model performance with different variable sizes](Image)
Oracle Advanced Analytics Database Architecture

Component of Oracle Database—SQL Functions

Oracle Database Enterprise Edition

Oracle Advanced Analytics

Native SQL Data Mining/Analytic Functions + High-performance R Integration for Scalable, Distributed, Parallel Execution
Fiserv
Combating Fraud

Objectives

- Build and implement Risk Mitigation Strategies for 2,500 US banks and Financial Institutions
- Fraud revention in online payments performed by organized sophisticated criminal groups
- When dealing with the hectic world of fraud, speed is the most important factor
- Hard to detect...target has low frequency (3 in 10,000)

Solution

- Oracle Advanced Analytics used by data analysts and deployed by DBA

“Oracle Advanced Analytics has a competitive advantage in terms of time savings, accuracy, cost, ease of use and deployment. When dealing with the hectic world of fraud, the speed to implement a new model is the most important factor. Systems with good algorithms and a fast turnaround have better ROI than systems with complex algorithms with long implementation times.”

- Miguel Barrera, Risk Manager, Julia Minkowski, Risk Analyst, Fiserv Inc.
What we learned...

- Complex Methods barely outperform simpler methods:
- Binning makes Trees and GLM almost as good as Ensemble or Gradient Methods
- Complex methods are hard to implement and require investments in infrastructure
- The current model building structure (SAS + Angoss) does not scale to grow with large volume
Accuracy + Agility vs. Cost to Deploy

- Pick the best combination of:
  - Less days to deployment
  - High model accuracy
  - Lower Cost

<table>
<thead>
<tr>
<th>Application</th>
<th>Deploy (Days)</th>
<th>Accuracy</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS Server</td>
<td>3</td>
<td>0.92</td>
<td>x5</td>
</tr>
<tr>
<td>ODM</td>
<td>1</td>
<td>0.90</td>
<td>1</td>
</tr>
<tr>
<td>SAS Base</td>
<td>15</td>
<td>0.83</td>
<td>30%</td>
</tr>
<tr>
<td>Angoss</td>
<td>12</td>
<td>0.85</td>
<td>10%</td>
</tr>
</tbody>
</table>
Predictive Analytics & Data Mining

Typical Use Cases

• Targeting the right customer with the right offer
• How is a customer likely to respond to an offer?
• Finding the most profitable growth opportunities
• Finding and preventing customer churn
• Maximizing cross-business impact
• Security and suspicious activity detection
• Understanding sentiments in customer conversations
• Reducing medical errors & improving quality of health
• Understanding influencers in social networks
What is Data Mining?

Automatically sifting through large amounts of data to find previously hidden patterns, discover valuable new insights and make predictions

- Identify most important factor (Attribute Importance)
- Predict customer behavior (Classification)
- Predict or estimate a value (Regression)
- Find profiles of targeted people or items (Decision Trees)
- Segment a population (Clustering)
- Find fraudulent or “rare events” (Anomaly Detection)
- Determine co-occurring items in a “baskets” (Associations)
Data Mining Provides Better Information, Valuable Insights and Predictions

Cell Phone Churners vs. Loyal Customers

Segment #1
IF CUST_MO > 14 AND INCOME < $90K, THEN Prediction = Cell Phone Churner,
Confidence = 100%
Support = 8/39

Segment #3
IF CUST_MO > 7 AND INCOME < $175K, THEN Prediction = Cell Phone Churner,
Confidence = 83%
Support = 6/39

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff
Oracle Advanced Analytics—**Best Practices**

Nothing is Different; Everything is Different

1. Start with a Business Problem Statement
2. Don’t Move the Data
3. Assemble the “Right Data” for the Problem
4. Create New Derived Variables
5. Be Creative in Analytical Methodologies
6. Quickly Transform “Data” to “Actionable Insights”
7. Automate and Deploy Enterprise-wide
Oracle Data Miner “Workflow” GUI for Data Analysts

**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
  - Save and share analytical workflows
  - Generate SQL scripts for deployment
Predicting Behavior
Identify “ Likely Behavior” and their Profiles

Consider:
- Demographics
- Past purchases
- Recent purchases
- Customer comments & tweets
Start with a Business Problem Statement

Common Examples

• Predict employees that voluntarily churn
• Predict customers that are likely to churn
• Target “best” customers
• Find items that will help me sell more most profitable items
• What is a specific customer most likely to purchase next?
• Who are my “best customers”?
• How can I combat fraud?
• I’ve got all this data; can you “mine” it and find useful insights?
Start with a Business Problem Statement

Clearly Define Problem

“If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions.”

— Albert Einstein
## Be Specific in Problem Statement

### Poorly Defined

- Predict employees that leave
- Predict customers that churn
- Target “best” customers
- How can I make more $$?
- Which customers are likely to buy?
- Who are my “best customers”?
- How can I combat fraud?
# Be Specific in Problem Statement

<table>
<thead>
<tr>
<th>Poorly Defined</th>
<th>Better</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict employees that leave</td>
<td>• Based on past employees that voluntarily left:</td>
</tr>
<tr>
<td></td>
<td>• Create New Attribute $\text{EmplTurnover} \rightarrow 0/1$</td>
</tr>
<tr>
<td>Predict customers that churn</td>
<td>• Based on past customers that have churned:</td>
</tr>
<tr>
<td></td>
<td>• Create New Attribute $\text{Churn} \rightarrow \text{YES/NO}$</td>
</tr>
<tr>
<td>Target “best” customers</td>
<td>• Recency, Frequency Monetary (RFM) Analysis</td>
</tr>
<tr>
<td></td>
<td>• Specific Dollar Amount over Time Window:</td>
</tr>
<tr>
<td></td>
<td>• Who has spent $500+ in most recent 18 months</td>
</tr>
<tr>
<td>How can I make more $$?</td>
<td>• What helps me sell soft drinks &amp; coffee?</td>
</tr>
<tr>
<td>Which customers are likely to buy?</td>
<td>• How much is each customer likely to spend?</td>
</tr>
<tr>
<td>Who are my “best customers”?</td>
<td>• What descriptive “rules” describe “best customers”?</td>
</tr>
<tr>
<td>How can I combat fraud?</td>
<td>• Which transactions are the most anomalous?</td>
</tr>
<tr>
<td></td>
<td>• Then roll-up to physician, claimant, employee, etc.</td>
</tr>
</tbody>
</table>
### Be Specific in Problem Statement

<table>
<thead>
<tr>
<th>Poorly Defined</th>
<th>Better</th>
<th>Data Mining Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict employees that leave</td>
<td>• Based on past employees that voluntarily left:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Create New Attribute \textbf{EmplTurnover → 0/1}</td>
<td></td>
</tr>
<tr>
<td>Predict customers that churn</td>
<td>• Based on past customers that have churned:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Create New Attribute \textbf{Churn → YES/NO}</td>
<td></td>
</tr>
<tr>
<td>Target “best” customers</td>
<td>• Recency, Frequency Monetary (RFM) Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Specific Dollar Amount over Time Window:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Who has spent $500+ in most recent 18 months</td>
<td></td>
</tr>
<tr>
<td>How can I make more $$?</td>
<td>• What helps me sell soft drinks &amp; coffee?</td>
<td></td>
</tr>
<tr>
<td>Which customers are likely to buy?</td>
<td>• How much is each customer likely to spend?</td>
<td></td>
</tr>
<tr>
<td>Who are my “best customers”?</td>
<td>• What descriptive “rules” describe “best customers”?</td>
<td></td>
</tr>
<tr>
<td>How can I combat fraud?</td>
<td>• Which transactions are the most anomalous?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Then roll-up to physician, claimant, employee, etc.</td>
<td></td>
</tr>
</tbody>
</table>
## Oracle Advanced Analytics

### In-Database Data Mining Algorithms—SQL & R & GUI Access

<table>
<thead>
<tr>
<th>Function</th>
<th>Algorithms</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
<td>Popular / Rules / transparency</td>
</tr>
<tr>
<td></td>
<td>Naïve Bayes</td>
<td>Embedded app</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machines (SVM)</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Linear Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine (SVM)</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Anomaly Detection</strong></td>
<td>One Class SVM</td>
<td>Unknown fraud cases or anomalies</td>
</tr>
<tr>
<td><strong>Attribute Importance</strong></td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction, Reduce data noise</td>
</tr>
<tr>
<td></td>
<td>Principal Components Analysis (PCA)</td>
<td></td>
</tr>
<tr>
<td><strong>Association Rules</strong></td>
<td>Apriori</td>
<td>Market basket analysis / Next Best Offer</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>Hierarchical k-Means</td>
<td>Product grouping / Text mining</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td></td>
<td>Expectation-Maximization Clustering (EM)</td>
<td></td>
</tr>
<tr>
<td><strong>Feature Extraction</strong></td>
<td>Nonnegative Matrix Factorization (NMF)</td>
<td>Text analysis / Feature reduction</td>
</tr>
<tr>
<td></td>
<td>Singular Value Decomposition (SVD)</td>
<td></td>
</tr>
</tbody>
</table>
In-Database Advanced Analytics

Independent Samples T-Test

- Query compares the mean of AMOUNT_SOLD between MEN and WOMEN Grouped By CUST_INCOME_LEVEL ranges
- Returns observed t value and its related two-sided significance (<.05 = significant)

```sql
SELECT substr(cust_income_level,1,22) income_level,
       avg(decode(cust_gender,'M',amount_sold,null)) sold_to_men,
       avg(decode(cust_gender,'F',amount_sold,null)) sold_to_women,
       stats_t_test_indep(cust_gender, amount_sold, 'STATISTIC','F') t_observed,
       stats_t_test_indep(cust_gender, amount_sold) two_sided_p_value
FROM sh.customers c, sh.sales s
WHERE c.cust_id=s.cust_id
GROUP BY rollup(cust_income_level)
ORDER BY 1;
```
R—Widely Popular

R is a statistics language similar to Base SAS or SPSS statistics

R environment

• Strengths
  – Powerful & Extensible
  – Graphical & Extensive statistics
  – Free—open source

• Challenges
  – Memory constrained
  – Single threaded
  – Outer loop—slows down process
  – Not industrial strength
Oracle Advanced Analytics
Oracle R Enterprise Compute Engines

R Engine
Oracle R Enterprise packages
Other R packages
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

Other R packages
Oracle R Enterprise Compute Engines

• 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

Oracle Database
SQL
User tables
Results

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 R Enterprise packages

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.
R Graphics Direct Access to Database Data

R> boxplot(split(CARSTATS$mpg, CARSTATS$model.year), col = "green")

MPG increases over time
Oracle Advanced Analytics

OAA/ORACLE R ENTERPRISE
QUICK DEMO
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
  – SQL Patterns
  – 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

* included in every Oracle Database
Data Mining When Lack Examples
Better Information, Valuable Insights and Predictions

Cell Phone Fraud vs. Loyal Customers

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff
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”
Fraud Prediction Demo

Automated In-DB Analytical Methodology

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;

<table>
<thead>
<tr>
<th>POLICYNUMBER</th>
<th>PERCENT_FRAUD</th>
<th>RNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>6532</td>
<td>64.78</td>
<td>1</td>
</tr>
<tr>
<td>2749</td>
<td>64.17</td>
<td>2</td>
</tr>
<tr>
<td>3440</td>
<td>63.22</td>
<td>3</td>
</tr>
<tr>
<td>654</td>
<td>63.1</td>
<td>4</td>
</tr>
<tr>
<td>12650</td>
<td>62.36</td>
<td>5</td>
</tr>
</tbody>
</table>

Automated Monthly “Application”!

Just add:
Create View CLAIMS2_30
As
Select * from CLAIMS2
Where mydate > SYSDATE – 30

Time measure: set timing on;
Retail

Market Basket Analysis

Find market baskets, product bundles, and next-likely products
Retail

Market Basket Analysis

• Perform market basket analysis in-database
• Find All “A→B rules”
• Sort by confidence
• Filter out recommendations that already in the customer’s shopping cart
• Finally, query the top 3 recommendations based on the order of highest confidence and support

```
SELECT rownum AS rank, consequent AS recommendation FROM 
  ( SELECT 
    cons_pred.attribute_subname consequent, 
    max(AR.rule_support) max_support, 
    max(AR.rule_confidence) max_confidence 
  FROM TABLE ( 
    DBMS_DATA_MINING.GET_ASSOCIATION_RULES ( 
      'AR_RECOMMENDATION', 10, NULL, 0.5, 0.01, 2, 1, 
      ORA_MINING_VARCHAR2_NT ( 
        'RULE_CONFIDENCE DESC', 'RULE_SUPPORT DESC'), 
      DM_ITEMS(DM_ITEM('PROD_NAME', 'Comic Book Heroes', NULL, NULL), 
      DM_ITEM('PROD_NAME', 'Martial Arts Champions', NULL, NULL), 
      NULL, 1)) AR, 
      TABLE(AR.consequent) cons_pred 
    WHERE cons_pred.attribute_subname NOT IN ('Comic Book Heroes', 'Martial Arts Champions') 
    GROUP BY cons_pred.attribute_subname 
    ORDER BY max_confidence DESC, max_support DESC 
  ) WHERE rownum <=3;
```

RANK RECOMMENDATION

---------------------
1 Endurance Racing
2 128MB Memory Card
3 Xtend Memory
Accelerates Complex Segmentation Queries from Weeks to Minutes—Gains Competitive Advantage

Objectives

- World’s leading customer-science company
- Accelerate analytic capabilities to near real time using Oracle Advanced Analytics and third-party tools, enabling analysis of unstructured big data from emerging sources, like smart phones

Solution

- Accelerated segmentation and customer-loyalty analysis from one week to just four hours—enabling the company to deliver more timely information & finer-grained analysis
- Generated more accurate business insights and marketing recommendations with the ability to analyze 100% of data—including years of historical data—instead of just a small sample

“Improved analysts’ productivity and focus as they can now run queries and complete analysis without having to wait hours or days for a query to process”

“Improved accuracy of marketing recommendations by analyzing larger sample sizes and predicting the market’s reception to new product ideas and strategies”

– dunnhumby Oracle Customer Snapshot
Oracle Advanced Analytics

More Details

- On-the-fly, single record apply with new data (e.g. from call center)

```
SELECT prediction_probability(CLAS_DT_4_15, 'Yes'
    USING 7800 as bank_funds, 125 as checking_amount, 20 as credit_balance, 55 as age, 'Married' as marital_status,
    250 as MONEY_MONTHLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```

Likelihood to respond:

Query Result

<table>
<thead>
<tr>
<th>PREDICTION_PROB...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8382936507936...</td>
</tr>
</tbody>
</table>

All Rows Fetched: 1 in 0 seconds
Fusion HCM Predictive Workforce

Predictive Analytics Applications

Fusion Human Capital Management Powered by OAA

- Oracle Advanced Analytics factory-installed predictive analytics
- Employees likely to leave and predicted performance
- Top reasons, expected behavior
- Real-time "What if?" analysis
Fusion HCM Predictive Workforce

Predictive Analytics Applications

Fusion Human Capital Management Powered by OAA

- Oracle Advanced Analytics factory-installed predictive analytics
- Employees likely to leave and predicted performance
- Top reasons, expected behavior
- Real-time "What if?" analysis
Oracle Communications Industry Data Model

Predictive Analytics Applications

- Enterprise wide data model for communications industry
  - Over 1,500 tables and 30,000 columns
  - Over 1,000 industry measures and KPIs
  - TMF SID conformance aligned
- Prebuilt mining models, OLAP cubes and sample reports
- Automatic data movement across layers
- Easily extensible and customizable
- Usable within any source application
Oracle Communications Industry Data Model

Predictive Analytics Applications

Pre-Built Predictive Models

• Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics
• OAA’s clustering and predictions available in-DB for OBIEE
• Automatic Customer Segmentation, Churn Predictions, and Sentiment Analysis

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.
Oracle Communications Data Model

Pre-Built Data Mining Models

1. Prepaid Churn Prediction
2. Postpaid Churn Prediction
3. Customer Profiling
4. Targeted Promotion
5. Customer Lifetime Value
6. Customer Lifetime Survival Value
7. Customer Sentiment
Oracle Communications Data Model

Pre-Built Prepaid Churn Prediction Data Mining Models

- Prepaid Churn Prediction Definition
  - Customer is recognized as a churner when he stop using any product from the operator
- Sample Input Attributes Used in Model
  - 170 attributes used in total for prepaid churn model

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCPT_NWSLTR_IND</td>
<td>Indicates whether customer accepts News Letter</td>
</tr>
<tr>
<td>BRDBND_IND</td>
<td>Indicates whether Customer has Broadband connection</td>
</tr>
<tr>
<td>CAR_DRV_LICNS_IND</td>
<td>Indicates whether customer has driver's license</td>
</tr>
<tr>
<td>CAR_TYP_CD</td>
<td>Car Type Code</td>
</tr>
<tr>
<td>CHRN_IND</td>
<td>Indicates whether a customer is a Churner or Non-churner</td>
</tr>
<tr>
<td>CMPLNT_CNT_LAST_3MO</td>
<td>Number of complaints made by customer in last 3 months</td>
</tr>
<tr>
<td>CMPLNT_CNT_LAST_MO</td>
<td>Number of complaints made by customer in this month</td>
</tr>
<tr>
<td>CMPLNT_CNT_LFTM</td>
<td>Number of complaints made by customer in his/her life span</td>
</tr>
<tr>
<td>CMPLNT_CNT_RPT</td>
<td>Customer Credit Category</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Customer Revenue Band Code</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Days between first payment and first recharge</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Days between payment and first use</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Number of dropped calls in last 3 months</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Number of dropped calls this month</td>
</tr>
<tr>
<td>CMPLNT_CNT_RX</td>
<td>Number of dropped calls in customer life span</td>
</tr>
<tr>
<td>DWLNG_OWNER</td>
<td>Dwelling Owner</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Dwelling Status</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Dwelling Size</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Dwelling Tenure</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Data downloaded in KBs in last 3 months</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Data downloaded in KBs in last 1 month</td>
</tr>
<tr>
<td>DWLNG_STA</td>
<td>Data downloaded in KBs in lifetime</td>
</tr>
<tr>
<td>ETHNCTY</td>
<td>Customer Ethnicity</td>
</tr>
<tr>
<td>GNDR_CD</td>
<td>Individual Customer Gender Code</td>
</tr>
<tr>
<td>HH_SZ</td>
<td>Household Size</td>
</tr>
<tr>
<td>HH_SZ</td>
<td>Number of hangup calls in last 3 months</td>
</tr>
<tr>
<td>HH_SZ</td>
<td>Number of hangup calls this month</td>
</tr>
<tr>
<td>HH_SZ</td>
<td>Number of offnet calls in last 1 month</td>
</tr>
<tr>
<td>HH_SZ</td>
<td>Indicates whether Customer has Pay TV connection</td>
</tr>
</tbody>
</table>
Integrated with OCDM, OBIEE, and leverages Oracle Data Mining with specialized SNA code

Identification of social network communities from CDR data

Predictive scores for churn and influence at a node level, as well as potential revenue/value at risk

User interface targeted at business users and flexible ad-hoc reporting
Integrated Business Intelligence
Enhance Dashboards with Predictions and Data Mining Insights

• In-database predictive models “mine” customer data and predict their behavior

• OBIEE’s integrated spatial mapping shows location

• All OAA results and predictions available in Database via OBIEE Admin to enhance dashboards

Customer “most likely” to be HIGH and VERY HIGH value customer in the future
Integrated Business Intelligence
Enhance Dashboards with Predictions and Data Mining Insights

• In-database predictive models “mine” customer data and predict their behavior
• OBIEE’s integrated spatial mapping shows location
• All OAA results and predictions available in Database via OBIEE Admin to enhance dashboards

Oracle BI EE defines results for end user presentation

Oracle Data Mining results available to Oracle BI EE administrators
Oracle Advanced Analytics Database Option

Oracle Data Miner 4.0 Summary New Features

• Oracle Data Miner/SQLDEV 4.0 (for Oracle Database 11g and 12c)
  – New **Graph node** (box, scatter, bar, histograms)
  – **SQL Query node** + integration of R scripts
  – Automatic **SQL script generation** for deployment

• Oracle Advanced Analytics 12c features exposed in Oracle Data Miner
  – New SQL data mining algorithms/enhancements
    • Expectation Maximization clustering algorithm
    • PCA & Singular Vector Decomposition algorithms
    • Improved/automated Text Mining, Prediction Details and other algorithm improvements
  – Predictive SQL Queries—automatic build, apply within SQL query
New Features

- **Graph node**
  - Scatter, line, bar, box plots, histograms
  - Group_by supported
New Features

• **SQL Query** node
  - Allows any form of query/transformation/statistics within an ODM’r work flow
  - Use SQL anywhere to handle special/unique data manipulation use cases
    • Recency, Frequency, Monetary (RFM)
    • SQL Window functions for e.g. moving average of $$ checks written past 3 months vs. past 3 days
  - Allows integration of R Scripts
SQL Developer/Oracle Data Miner 4.0

New Features

- **SQL Script Generation**
  - Deploy entire methodology as a SQL script
  - Immediate deployment of data analyst’s methodologies
SQL Developer/Oracle Data Miner 4.0

New Features

• **SQL Query node**
  – Allows integration of R Scripts
SQL Developer/Oracle Data Miner 4.0

New Features

- **SQL Query node**
  - Allows integration of R Scripts

```
SELECT VALUE FROM TABLE
{
  rTableEval:
  cursor (select * from "Filter NUMBER Column_B10010"),
  NULL,
  'XML',
  'R RANDOM FOREST MODEL BUILD')
}
```
12c New Features

New Server Functionality

• 3 New Oracle Data Mining SQL functions algorithms
  – Expectation Maximization (EM) Clustering
    • New Clustering Technique
      – Probabilistic clustering algorithm that creates a density model of the data
      – Improved approach for data originating in different domains (for example, sales transactions and customer demographics, or structured data and text or other unstructured data)
      – Automatically determines the optimal number of clusters needed to model the data.
  – Principal Components Analysis (PCA)
    • Data Reduction & improved modeling capability
      – Based on SVD, powerful feature extraction method use orthogonal linear projections to capture the underlying variance of the data
  – Singular Value Decomposition (SVD)
    • Big data “workhorse” technique for matrix operations
      – Scales well to very large data sizes (both rows and attributes) for very large numerical data sets (e.g. sensor data, text, etc.)
12c New Features

New Server Functionality

• Text Mining Support Enhancements
  – This enhancement greatly simplifies the data mining process (model build, deployment and scoring) when text data is present in the input:
    • Manual pre-processing of text data is no longer needed.
    • No text index needs to be created.
    • Additional data types are supported: CLOB, BLOB, BFILE.
    • Character data can be specified as either categorical values or text.
12c New Features

New Server Functionality

• Predictive Queries
  – Immediate build/apply of ODM models in SQL query
  • Classification & regression
    – Multi-target (nested) problems
  • Clustering query
  • Anomaly query
  • Feature extraction query

OAA automatically creates multiple anomaly detection models “Grouped_By” and “scores” by partition via powerful SQL query

Results/Predictions!
The Four Traps of Predictive Analytics

• The First Trap: Magical Thinking
  – The need to really understand what you want to decide using analytics before you develop them
    “If I had an hour to solve a problem I’d spend 55 minutes thinking about the problem and 5 minutes thinking about solutions.”
    — Albert Einstein

• The Second Trap: Starting at the Top
  – The need to begin with operational decisions not strategic ones
  – Predictive analytics works best for prompting decisions about operations, rather than initiating their use at the executive level.
  – Operational decisions, such as those in which companies choose a supplier or determine whether to extend credit, lend themselves well to predictive analytics
  – Companies also need to frame their predictive analytics around actions. “Don’t look at how good a customer is.
    • Look at, what action should I offer to a customer?
    • Should I change suppliers?”
The Four Traps of Predictive Analytics

• The Third Trap: Building Cottages, Not Factories
  – Need to industrialize analytics not treat it as a cottage industry. Creating analytic models that don’t scale. Analytics specialists are no more connected to the business than technology specialists.
  – Otherwise, analytics specialists are prone to create the equivalent of a cottage industry, where the models built apply to only one thing, or are too complex and expensive to be reused easily.
    • Netflix’s famous challenge, where it gave $1 million for a better algorithm to make movie recommendations. Its million-dollar model “was never deployed,” Taylor said. “They got a fabulous model, but ask them, and they will tell you that the resources weren’t available to use it. What they meant to fund was ‘a model that was more predictive that we can realistically deploy and run on our service in Earth time.’ They didn’t ask for that.”

• The Fourth Trap: Seeking Purified Data
  – Avoid being paralyzed by weakness in your data
    • “Garbage in, garbage out” is the cliché of data-haters everywhere. “It is not true that companies need good data to use predictive analytics,” Taylor said. “The techniques can be robust in the face of terrible data, because they were invented by people who had terrible data,” he noted.
  – Companies should start with the business decision they want to make, and then look for data that might help them predict outcomes.
OAA Links and Resources

• **Oracle Advanced Analytics Overview:**
  - Link to presentation—[Big Data Analytics using Oracle Advanced Analytics In-Database Option](#)
  - [OAA data sheet](#) on OTN
  - [Oracle Internal OAA Product Management Wiki and Workspace](#)

• **YouTube recorded OAA Presentations and Demos:**
  - [Oracle Advanced Analytics and Data Mining at the YouTube Movies](#) (6 + OAA “live” Demos on ODM’r 4.0 New Features, Retail, Fraud, Loyalty, Overview, etc.)

• **Getting Started:**
  - Link to [Getting Started w/ ODM blog entry](#)
  - Link to [New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course](#).
  - Link to [OAA/Oracle Data Mining 4.0 Oracle by Examples (free) Tutorials](#) on OTN
  - Take a [Free Test Drive of Oracle Advanced Analytics (Oracle Data Miner GUI) on the Amazon Cloud](#)
  - Link to [SQL Developer Days Virtual Event w/ downloadable VM of Oracle Database + ODM/ODMr and e-training for Hands on Labs](#)
  - Link to [OAA/Oracle R Enterprise (free) Tutorial Series](#) on OTN

• **Additional Resources:**
  - [Oracle Advanced Analytics Option on OTN page](#)
  - [OAA/Oracle Data Mining on OTN page](#), [ODM Documentation & ODM Blog](#)
  - [OAA/Oracle R Enterprise page on OTN page](#), [ORE Documentation & ORE Blog](#)
  - [Oracle SQL based Basic Statistical functions](#) on OTN
  - Business Intelligence, Warehousing & Analytics—[BIWA Summit’15, Jan 27-29, 2015](#) at Oracle HQ Conference Center
New book on Oracle Advanced Analytics available

Book available on Amazon
Predictive Analytics Using Oracle Data Miner: Develop for ODM in SQL & PL/SQL
Take a Test Drive!
Vlamis Software, Oracle Partner Offers FREE Test Drives on the Amazon Cloud

• Step 1—Fill out request
  – Go to http://www.vlamis.com/testdrive-registration/

• Step 2—Connect
  – Connect with Remote Desktop

• Step 3—Start Test Drive!
  – Oracle Database +
  – Oracle Advanced Analytics Option
  – SQL Developer/Oracle Data Miner GUI
  – Demo data for learning
  – Follow Tutorials
BIWA Summit
January 27-29, 2015
Oracle HQ Conference Center
www.biwasummit.org