Oracle Advanced Analytics
Make Big Data + Analytics Simple

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.
Agenda

• Big Data + Analytics phenomenon
• Oracle Advanced Analytics overview & features/benefits
  – GUI
  – SQL data mining functions
  – R integration
• Brief demos
• Big Data SQL
• Applications “powered by OAA”
• Getting started
Planning for Future
Growth of Data Exponentially Greater than Growth of Data Analysts!

Growth of Data vs. Growth of Data Analysts

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

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

http://www.delphianalytics.net/more-data-than-analysts-the-real-big-data-problem/
Analytics + Data Warehouse + Hadoop

• Platform Sprawl
  – More Duplicated Data
  – More Data Movement Latency
  – More Security challenges
  – More Duplicated Storage
  – More Duplicated Backups
  – More Duplicated Systems
  – More Space and Power
Vision

• Creating an Big Data + Analytic Platform for the Era of Big Data and the Cloud
  – Make Big Data + Analytics **Simple**
    • Any data size, on any computer infrastructure
    • Any variety of data, in any combination
  – Make Big Data + Analytics Deployment **Simple**
    • As a service, as a platform, as an application
Oracle Advanced Analytics Database Evolution

- Oracle acquires Thinking Machine Corp’s dev. team + “Darwin” data mining software
- 7 Data Mining “Partners”

1998 1999

- Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API
- Oracle Data Mining 10g & 10gR2 introduces SQL dm functions, 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI

2002 2004

- ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements
- Integration with “R” and introduction/addition of Oracle R Enterprise
- Product renamed “Oracle Advanced Analytics (ODM + ORE)”

2005 2008

- 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
- SQLDEV/Oracle Data Miner adds NN, Stepwise, scalable R algorithms

2011 2014

- Oracle Adv. Analytics for Hadoop Connector launched with scalable BDA algorithms
- Integration with “R” and introduction/addition of Oracle R Enterprise
- Product renamed “Oracle Advanced Analytics (ODM + ORE)”

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.
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
- Trilingual component of Oracle Database—SQL, SQLDev/ODMr GUI, R
- Scalable, parallel in-database execution
- Workflow GUI and IDEs
- Integrated component of Database
- Enables enterprise analytical applications
Oracle Advanced Analytics Database Option
Trilingual Component of Oracle Database—SQL, SQLDev/ODMr GUI, R

Key Features

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

Oracle Advanced Analytics

Traditional Analytics

- Data Import
- Data Mining
  - Model “Scoring”
- Data Prep. & Transformation
- Data Mining
  - Model Building
- Data Prep & Transformation
- Data Extraction

Savings

Secs, Mins or Hours

Hours, Days or Weeks

Data Extraction

Data Prep. & Transformation

Data Mining
  - Model Building

Data Mining
  - Model “Scoring”

Embeded Data Prep

Data Preparation

Model Building

Model “Scoring”

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.
Oracle Advanced Analytics Database Option
Trilingual Component of Oracle Database—SQL, SQLDev/ODMg GUI, R

Key Features

Lowest Total Cost of Ownership
- Eliminate data duplication
- Eliminate separate analytical servers
- Leverage investment in Oracle IT

Fastest way to deliver enterprise-wide predictive analytics
- Integrated GUI for Predictive Analytics
- Database scoring engine
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 Database Architecture
Trilingual Component of Oracle Database—SQL, SQLDev/ODMr GUI, R

Users
Data & Business Analysts: SQL Developer
R programmers: R Client
Business Analysts/Mgrs: OBIEE
Domain End Users: Applications

Platform
Oracle Database Enterprise Edition
Oracle Advanced Analytics
Native SQL Data Mining/Analytic Functions + High-performance
R Integration for Scalable, Distributed, Parallel Execution
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.

- Naïve Guess or Random
- Model with 20 variables
- Model with 75 variables
- Model with 250 variables
# 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>
Oracle Advanced Analytics Database Option

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 & Spatial Data
  – All OAA algorithms support transactional data (i.e. purchase transactions, repeated measures over time, distances from location, time spent in area A, B, C, etc.)

• 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 free in every Oracle Database
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
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

Predicting Behavior
Identify “Likely Behavior” and their Profiles

SQL Joins and arbitrary SQL transforms & queries – power of SQL

Consider:
• Demographics
• Past purchases
• Recent purchases
• Customer comments & tweets

Unstructured data also mined by algorithms

Transactional POS data

Generates SQL scripts for deployment

Inline predictive model to augment input data

Aggregated POS data

POS: Sales data

Customers

Credit score model

Predicted credit rating

Join

360 degree view of customers

Filter Columns

Multiple Predictive models

Most Likely customers

Demographics and comments

Customer sentiment data

Clust Build
Oracle Advanced Analytics

Brief Demos
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
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>63.10</td>
<td>4</td>
</tr>
<tr>
<td>12650</td>
<td>62.36</td>
<td>5</td>
</tr>
</tbody>
</table>
Oracle Advanced Analytics

More Details

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

```
SELECT prediction_probability(CLAS_DT_1_2, 'Yes'
    USING 7800 as bank_funds, 125 as checking_amount, 20 as credit_balance, 55 as age, 'Married' as marital_status, 250 as MONEY_MONLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```

Likelihood to respond:

- Query Result
  - All Rows Fetched: 1 in 0 seconds
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”
Tax Noncompliance Audit Selection

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

OAA/Oracle R Enterprise (R integration)
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

User R Engine on desktop
- R-SQL Transparency Framework overloads R functions for scalable in-database execution
- Function overload for data transforms, statistical functions and advanced analytics
- Interactive display of graphical results and flow control as in standard R
- Submit user-defined R functions for execution at database server under control of Oracle 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 production deployment and automated execution of R scripts

Oracle R Enterprise Compute Engines

1. R Engine
2. Oracle Database
3. Other R packages

Oracle R Enterprise packages

Oracle R Enterprise Compute Engines

R Engine

Oracle Database

User tables

Results

SQL

Oracle R Enterprise packages

Other R packages

R Engine

Oracle R Enterprise packages

Other R packages

1

2

3

Open Source

R

Oracle R Enterprise Compute Engines

Results

Oracle R Enterprise Compute Engines

Results

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
R: Transparency through function overloading

Invoke in-database aggregation function

```r
aggdata <- aggregate(ONTIME_S$DEST,
  by = list(ONTIME_S$DEST),
  FUN = length)

class(aggdata)
[1] "ore.frame"
attr("package")
[1] "OREbase"

head(aggdata)
Group.1 x
 1 ABE 237
 2 ABI 34
 3 ABQ 1357
 4 ABY 10
 5 ACK 3
 6 ACT 33
```

Oracle SQL

```sql
select DEST, count(*)
from ONTIME_S
group by DEST
```

Oracle Database

In-db Stats

```
ONTIME_S
```

Database Server
R: Transparency through function overloading

Invoke in-database Data Mining model (Support Vector Machine)

```r
> svm_mod <- ore.odmSVM(BUY~INCOME+YRS_CUST+MARITAL_STATUS, data=CUST, 
>                         "classification", kernel="linear")

> summary(svm_mod)

Call:
ore.odmSVM(formula = BUY ~ INCOME + YRS_CUST + MARITAL_STATUS, data = CUST, 
            type = "classification", kernel.function = "linear")

Settings:

value
prep.auto on
active.learning al.enable
complexity.factor 46.044899
conv.tolerance 1e-04
kernel.function linear

Coefficients:

                  class variable  value     estimate
  1     0          INCOME 5.204561e-05
  2     0 MARITAL_STATUS M 4.531359e-05
  3     0 MARITAL_STATUS S 4.531359e-05
  4     0          YRS_CUST 1.264948e-04
  5     0     (Intercept) 9.999269e-01
  6     1          INCOME 2.032340e-05
  7     1 MARITAL_STATUS M 2.636552e-06
  8     1 MARITAL_STATUS S 2.636555e-06
  9     1          YRS_CUST 1.588211e-04
 10     1     (Intercept) 9.999324e-01
```
Oracle Advanced Analytics for Hadoop

Predictive algorithms that execute in a parallel/distributed manner on Hadoop with data in HDFS
Oracle R Advanced Analytics for Hadoop

• ORAAH = Oracle R Advanced Analytics for Hadoop, part of Big Data Software Connectors Suite (Oracle Big Data Appliance Option)

• ORAAH transparency layer enables certain overloaded R functions to operate on Hive tables using R syntax and behavior (transparently translating R to HiveQL)

• R interface for manipulating HDFS data and writing mapper and reducer functions in R – where you can leverage open source CRAN packages – and invoke those Hadoop jobs from R

• Provides a range of predictive algorithms that execute on the Hadoop cluster with data in HDFS in a parallel/distributed manner.
Oracle R Advanced Analytics for Hadoop
Collection of R packages that provide:

• Interfaces with Apache Hive tables, the Apache Hadoop infrastructure, local R environment, and Oracle database tables
• Predictive analytic techniques, written in R or Java as Hadoop MapReduce jobs, that can be applied to data in HDFS files
• Install and load package as you would any R package to perform e.g.:
  – Access and transform HDFS data using a Hive-enabled transparency layer
  – Use the R language for writing mappers and reducers
  – Copy data between R memory, the local file system, HDFS, Hive, and Oracle databases
  – Schedule R programs to execute as Hadoop MapReduce jobs & return the results to those locations
• To use Oracle R Advanced Analytics for Hadoop, user should be familiar with MapReduce programming, R programming, and statistical methods.
### ORAAH in-Hadoop MR Functions

**Current release**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>orch.cor</td>
<td>Generates a correlation matrix with a Pearson's correlation coefficients.</td>
</tr>
<tr>
<td>orch.cov</td>
<td>Generates a covariance matrix.</td>
</tr>
<tr>
<td>orch.getXlevels</td>
<td>Creates a list of factor levels that can be used in the xlev argument of a model.matrix call. It is equivalent to the .getXlevels function in the stats package.</td>
</tr>
<tr>
<td>orch.glm</td>
<td>Fits and uses generalized linear models on data stored in HDFS.</td>
</tr>
<tr>
<td>orch.kmeans</td>
<td>Perform k-means clustering on a data matrix that is stored as a file in HDFS.</td>
</tr>
<tr>
<td>orch.lm</td>
<td>Fits a linear model using tall-and-skinny QR (TSQR) factorization and parallel distribution. The function computes the same statistical parameters as the Oracle R Enterprise ore.lm function.</td>
</tr>
<tr>
<td>orch.lmf</td>
<td>Fits a low rank matrix factorization model using either the jellyfish algorithm or the Mahout alternating least squares with weighted regularization (ALS-WR) algorithm.</td>
</tr>
</tbody>
</table>
## ORAAH in-Hadoop MR Functions

### Current release

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>orch.neural</code></td>
<td>Provides a neural network to model complex, nonlinear relationships between inputs and outputs, or to find patterns in the data.</td>
</tr>
<tr>
<td><code>orch.nmf</code></td>
<td>Provides the main entry point to create a nonnegative matrix factorization model using the jellyfish algorithm. This function can work on much larger data sets than the R NMF package, because the input does not need to fit into memory.</td>
</tr>
<tr>
<td><code>orch.nmf.NMFalgo</code></td>
<td>Plugs in to the R NMF package framework as a custom algorithm. This function is used for benchmark testing.</td>
</tr>
<tr>
<td><code>orch.princomp</code></td>
<td>Analyzes the performance of principal component.</td>
</tr>
<tr>
<td><code>orch.recommend</code></td>
<td>Computes the top $n$ items to be recommended for each user that has predicted ratings based on the input <code>orch.mahout.lmf.asl</code> model.</td>
</tr>
<tr>
<td><code>orch.sample</code></td>
<td>Provides the reservoir sampling.</td>
</tr>
<tr>
<td><code>orch.scale</code></td>
<td>Performs scaling.</td>
</tr>
</tbody>
</table>
Big Data SQL

Push down SQL predicts to storage layers
What gives Exadata **extreme** performance?
Data Analytics Challenge
Separate silos with separate data access interfaces

{APIs}  SQL  {MapReduce}

NoSQL  ORACLE
What customers want: Oracle Big Data SQL
Rich, comprehensive SQL access to all enterprise data

The Power of Oracle SQL
- Wide variety of ‘Big Data’ types
  - Structured data
    - Numeric, string, date, ...
  - Unstructured data
    - LOBs, Text, XML, JSON, Spatial, Graph, Multimedia
- Rich SQL Analytic Functions
  - Ranking, Windowing, LAG/LEAD, Aggregate, Pattern Matching, Cross Tabs, Statistical, Linear Regression, Correlations, Hypothesis Testing, Distribution Fitting, ...
Introducing Oracle Big Data SQL
Massively Parallel SQL Query across Oracle, Hadoop and NoSQL

Offload Query to Data Nodes

Offload Query to Exadata Storage Servers

Small data subset quickly returned

Hadoop & NoSQL
Oracle Database 12c

data subset
Manage and **Analyze** All Data—SQL & Oracle Big Data SQL

- **Oracle Big Data Appliance**
  - Store JSON data unconverted in Hadoop

- **Oracle Database 12c**
  - Store business-critical data in Oracle
  - Data analyzed via SQL or R

SQL
Oracle Advanced Analytics
Applications Integration + OBIEE Integration
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
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
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 Data Model

Pre-Built Data Mining Models

1. Churn Prediction
2. Customer Profiling
3. Customer Churn Factor
4. Cross-Sell Opportunity
5. Customer Life Time Value
6. Customer Sentiment
7. Customer Life Time Value
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>CRDT_CTGRY_KEY</td>
<td>Customer Credit Category</td>
</tr>
<tr>
<td>CUST_RVN_BND_CD</td>
<td>Customer Revenue Band Code</td>
</tr>
<tr>
<td>DAYS_BFR_FIRST_RCC</td>
<td>Days between first payment and first recharge</td>
</tr>
<tr>
<td>DAYS_BFR_FIRST_USE</td>
<td>Days between payment and first use</td>
</tr>
<tr>
<td>DRPD_CALLS_CNT_LAST_3MO</td>
<td>Number of dropped calls in last 3 months</td>
</tr>
<tr>
<td>DRPD_CALLS_CNT_LAST_MO</td>
<td>Number of dropped calls this month</td>
</tr>
<tr>
<td>DRPD_CALLS_CNT_LFTM</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_SZ</td>
<td>Dwelling Size</td>
</tr>
<tr>
<td>DWLNG_TENR</td>
<td>Dwelling Tenure</td>
</tr>
<tr>
<td>DWNLD_DATA_LAST_3MO</td>
<td>Data downloaded in KBs in last 3 months</td>
</tr>
<tr>
<td>DWNLD_DATA_LAST_MO</td>
<td>Data downloaded in KBs in last 1 month</td>
</tr>
<tr>
<td>DWNLD_DATA_LFTM</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>HNGUP_CALLS_CNT_LAST_3MO</td>
<td>Number of hangup calls in last 3 months</td>
</tr>
<tr>
<td>HNGUP_CALLS_CNT_LAST_MO</td>
<td>Number of hangup calls this month</td>
</tr>
<tr>
<td>MMS_CNT_LAST_MO</td>
<td>MMSs sent in last 1 month</td>
</tr>
<tr>
<td>OFFNET_CALLS_LAST_MO</td>
<td>Number of offnet calls in last 1 month</td>
</tr>
<tr>
<td>PAY_TV_IND</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
12c New Features
Oracle Advanced Analytics Database Option

Oracle Data Miner 4.X Summary New Features

• Oracle Data Miner/SQLDEV 4.1 EA2 (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
  – **JSON Query node** to mine Big Data external tables

• 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
SQL Developer/Oracle Data Miner 4.0

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
New Features

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

• Database/Data Mining Parallelism On/Off Control
  – Allows users to take full advantage of Oracle parallelism/scalability on an Oracle Data Miner node by node basis
  • Default is “Off”
  – Important for large Oracle Database & Oracle Exadata shops
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 problems
• Clustering query
• Anomaly query
• Feature extraction query

Select
cust_income_level, cust_id,
round(probanom,2) probanom, round(pctrank,3)*100 pctrank from ( select
cust_id, cust_income_level, probanom,
percent_rank() over (partition by cust_income_level order by probanom desc) pctrank
from ( select
  cust_id, cust_income_level,
prediction_probability(of anomaly, 0 using *)
  over (partition by cust_income_level) probanom
  from customers
) )
where pctrank <= .05
order by cust_income_level, probanom desc;

OAA automatically creates multiple anomaly detection models “Grouped_By” and “scores” by partition via powerful SQL query
12c New Features

New Server Functionality

• Predictive Queries
  – Immediate build/apply of ODM models in SQL query
    • Classification & regression
      – Multi-target 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!
New Features

• JSON Query node

JSON Query node extracts BDA data via External Tables and parses out JSON data type and assembles data for data mining.
New Features

- **Oracle Data Miner Workflow API to Manage, Schedule and Run Workflows**
  - PL/SQL APIs to enable applications to execute workflows immediately or schedule them
  - Oracle Scheduler for scheduling functionality
  - ODMr repository views can be queried for project and workflow information
  - Applications can monitor workflow execution and query generated results

```sql
CONNECT DMUSER/DMUSER
SET SERVEROUTPUT ON
DECLARE
    v_jobId VARCHAR2(30) := NULL;
    v_status VARCHAR2(30) := NULL;
    v_projectName VARCHAR2(30) := 'Project';
    v_workflow_name VARCHAR2(30) := 'build_workflow';
    v_node VARCHAR2(30) := 'MODEL_COEFFICIENTS';
    v_run_mode VARCHAR2(30) := ODMRSYS.ODMR_WORKFLOW.RERUN_NODE_PARENTS;
    v_failure NUMBER := 0;
    v_nodes ODMRSYS.ODMR_OBJECT_NAMES := ODMRSYS.ODMR_OBJECT_NAMES();
BEGIN
    v_nodes.extend();
    v_nodes(v_nodes.count) := v_node;
    v_jobId := ODMRSYS.ODMR_WORKFLOW.WF_RUN(p_project_name => v_projectName,
                                              p_workflow_name => v_workflow_name,
                                              p_node_names => v_nodes,
                                              p_run_mode => v_run_mode,
                                              p_start_date => '31-DEC-14 12.00.00 AM AMERICA/NEW_YORK',
                                              p_repeat_interval => 'FREQ=MONTHLY;BYMONTHDAY=-1',
                                              p_end_date => '31-DEC-15 12.00.00 AM AMERICA/NEW_YORK');
    DBMS_OUTPUT.PUT_LINE('Job: ' || v_jobId);
END;
```
Getting started
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
Click for BIWA Summit 2015 info
Thank you to all that made BIWA Summit 2015 which took place January 27-29, 2015 such a fantastic success! BIWA Summit 2015 presentations will be available soon to attendees.

Oracle BIWA Summit 2016
Including Oracle Spatial Summit 2016
In partnership with NoCOUG

January 26-28, 2016
Oracle Conference Center at Oracle HQ Campus, Redwood Shores, CA

BIWA Summit 2016 is being planned now. Send an email to Dan Vlamis at dvlamis@vlamis.com or Shyam Varan Nath at shyamvaran@gmail.com to get involved.

See the BIWA Summit 2015 home page for more information on BIWA Summit 2015, including presentations, agenda, etc.