Oracle’s Advanced Analytics
Making Big Data + Analytics Simple

Charlie Berger, MS Engineering, 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.
Predictive Analytics 101

• Data, data everywhere – explosive growth
• Growth of data exponentially greater than growth of data analysts!

The Useful Data GAP

Executive who feel they understand the impact data will have on their organizations

12%

Produce

Data

Use

Data

Machine Learning/Data Analysis platforms requirements:

• Be extremely powerful and handle large data volumes
• Be easy to learn
• Be highly automated & enable deployment

http://www.delphianalytics.net/more-data-than-analysts-the-real-big-data-problem/
Machine Learning/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

• Big Data + Machine Learning/Analytics Platform for the Era of Big Data and Cloud

  – Make Big Data + ML/Analytics Model Discovery _Simple_
    • Any data size, on any computer infrastructure—on-premise and/or cloud
    • Any variety of data (structured, unstructured, transactional, geospatial), in any combination

  – Make Big Data + ML/Analytics Model Deployment _Simple_
    • As a service, as a platform, as an application
    • On-premise and/or cloud
What is Machine Learning, Data Mining & Predictive Analytics?

Automatically sifting through large amounts of data to create models that 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)*
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
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Parallel, scalable data mining algorithms and R integration
- In-Database + Hadoop—Don’t move the data
- Data analysts, data scientists & developers
- Drag and drop workflow, R and SQL APIs
- Extends data management into powerful advanced/predictive analytics platform
- Enables enterprise predictive analytics deployment + applications
Oracle’s Advanced Analytics

Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Parallel, scalable data mining algorithms and R integration
- In-Database + Hadoop—Don’t move the data
- Data analysts, data scientists & developers
- Drag and drop workflow, R and SQL APIs
- Extends data management into powerful advanced/predictive analytics platform
- Enables enterprise predictive analytics deployment + applications

Don’t move data; Data is LARGE
Move the algorithms instead
Google “Oracle Advanced Analytics”

Oracle Advanced Analytics 12c delivers parallel in-database implementations of data mining algorithms and integration with open source R. Data analysts use Oracle Data Miner GUI and R to build and evaluate predictive models and leverage R packages and graphs. Application developers deploy Oracle Advanced Analytics models using SQL data mining functions and R. With the Oracle Advanced Analytics option, Oracle extends the Oracle Database to an analytic analytical environment that brings more data and data types, eliminates data movement, and provides security to anticipate customer behavior, detect patterns, and deliver actionable insights.

Oracle Advanced Analytics, a combination of Oracle Data Mining and Oracle R Enterprise, delivers predictive analytics, data mining, text mining, statistical analysis, advanced numerical computations and interactive graphics inside the database. It brings powerful computations to the database resulting in dramatic improvements in information discovery, scalability, security, and savings. Data analysts, data scientists, statistical programmers, application developers and DBAs can develop and automate sophisticated analytical methodologies inside the database and gain competitive advantage.

Oracle R Advanced Analytics for Hadoop (ORAAH) is one of the components in the Oracle Big Data Connections software suite, an option to the Oracle Big Data Appliance. ORAAH provides an R interface for managing data within ODI, using both HIVE transparency calculations and mapping and R-objects as data flows into machine learning algorithms.

The newly released Oracle R Advanced Analytics for Hadoop 2.0.0 includes ten new algorithm implementations that can take advantage of an Apache Spark cluster for a significant performance gain in model building and scoring time. These algorithms are a redesigned version of the Convex Layer Perceptron Neural Networks (convnet) and a brand new implementation of a Logistic Regression model (logitglm).
Oracle Advanced Analytics Database Evolution

- **1998**
  - Oracle acquires Thinking Machine Corp’s dev. team + “Darwin” data mining software

- **1999**
  - 7 Data Mining “Partners”

- **2002**
  - Oracle Data Mining 9.2i launched – 2 algorithms (NB and AR) via Java API

- **2004**
  - Oracle Data Mining 10gR2 SQL - 7 new SQL dm algorithms and new Oracle Data Miner “Classic” wizards driven GUI

- **2005**
  - SQL statistical functions introduced

- **2008**
  - ODM 11g & 11gR2 adds AutoDataPrep (ADP), text mining, perf. improvements

- **2011**
  - Integration with “R” and introduction/addition of Oracle R Enterprise

- **2014**
  - Product renamed “Oracle Advanced Analytics (ODM + ORE)
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide ML/Predictive Analytics

Major Benefits

- Data remains in Database & Hadoop
  - Model building and scoring occur in-database
  - Use R packages with data-parallel invocations
- Leverage investment in Oracle IT
  - Eliminate data duplication
  - Eliminate separate analytical servers
- Deliver enterprise-wide applications
  - GUI for ML/Predictive Analytics & code gen
  - R interface leverages database as HPC engine
Oracle’s Advanced Analytics (Machine Learning Platform)
Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

Information Producers

- R programmers
- Data & Business Analysts

Information Consumers

- Business Analysts/Mgrs
- Domain End Users

Users

Platform

- Hadoop
  - ORAAH
  - Parallel, distributed algorithms

Oracle Database Enterprise Edition

Oracle Advanced Analytics - Database Option

- SQL Data Mining, ML & Analytic Functions + R Integration
- for Scalable, Distributed, Parallel in-Database ML Execution

Oracle Cloud
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
You Can Think of Oracle Advanced Analytics Like This...

Traditional SQL

- “Human-driven” queries
- Domain expertise
- Any “rules” must be defined and managed

SQL Queries

- SELECT
- DISTINCT
- AGGREGATE
- WHERE
- AND OR
- GROUP BY
- ORDER BY
- RANK

SQL Statistical Functions - SQL & R

- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine/analyze

Statistical SQL “Verbs”

- MEAN, STDEV
- MEDIAN
- SUMMARY
- CORRELATE
- FIT
- COMPARE
- ANOVA

FREE!
In-Database Statistical Functions (SQL)

Independent Samples T-Test

• A/B offer testing
  – 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;
```
Correlation Functions

- The CORR_S and CORR_K functions support nonparametric or rank correlation (finding correlations between expressions that are ordinal scaled).
- Correlation coefficients take on a value ranging from -1 to 1, where:
  - 1 indicates a perfect relationship
  - -1 indicates a perfect inverse relationship
  - 0 indicates no relationship
- The following query determines whether there is a correlation between the AGE and WEIGHT of people, using Spearman's correlation:

```
select CORR_S(AGE, WEIGHT) coefficient,
       CORR_S(AGE, WEIGHT, 'TWO_SIDED_SIG') p_value,
       substr(TREATMENT_PLAN, 1,15) as TREATMENT_PLAN
from DMUSER.LYMPHOMA
GROUP BY TREATMENT_PLAN;
```
You Can Think of Oracle’s Advanced Analytics Like This…

**Traditional SQL**
- “Human-driven” queries
- Domain expertise
- Any “rules” must be defined and managed

**SQL Queries**
- SELECT
- DISTINCT
- AGGREGATE
- WHERE
- AND OR
- GROUP BY
- ORDER BY
- RANK

**Oracle Advanced Analytics - SQL & R**
- Automated knowledge discovery, model building and deployment
- Domain expertise to assemble the “right” data to mine/analyze

**Analytical SQL “Verbs”**
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Oracle Advanced Analytics Database Option

In-Database ML/Data Mining Algorithms*—SQL & R & GUI Access

**Classification**
- Decision Tree
- Logistic Regression (GLM)
- Naïve Bayes
- Support Vector Machine (SVM)
- Random Forest

**Regression**
- Multiple Regression (GLM)
- Support Vector Machine (SVM)
- Linear Model
- Generalized Linear Model
- Multi-Layer Neural Networks
- Stepwise Linear Regression

**Clustering**
- Hierarchical k-Means
- Orthogonal Partitioning Clustering
- Expectation-Maximization

**Attribute Importance**
- Minimum Description Length
- Unsupervised pair-wise KL div.

**Anomaly Detection**
- 1 Class Support Vector Machine

**Time Series**
- Single & Double Exp. Smoothing

**Predictive Queries**
- Clustering
- Regression
- Anomaly Detection
- Feature Extraction

**Feature Extraction & Creation**
- Nonnegative Matrix Factorization
- Principal Component Analysis
- Singular Value Decomposition

**Market Basket Analysis**
- Apriori – Association Rules

**Open Source R Algorithms**
- Ability to run any R package via Embedded R mode
  * supports partitioned models, text mining

---

* supports partitioned models, text mining

---
Oracle Advanced Analytics
How Oracle R Enterprise Compute Engines Work

1. **R-> SQL Transparency “Push-Down”**
   - R language for interaction with the database
   - R-SQL Transparency Framework overloads R functions for scalable in-database execution
   - Function overload for data selection, manipulation and transforms
   - 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

2. **In-Database Adv Analytical SQL Functions**
   - 15+ Powerful data mining algorithms (regression, clustering, AR, DT, etc.)
   - Run Oracle Data Mining SQL data mining functioning (ORE.odmSVM, ORE.odmDT, etc.)
   - Speak “R” but executes as proprietary in-database SQL functions—machine learning algorithms and statistical functions
   - Leverage database strengths: SQL parallelism, scale to large datasets, security
   - Access big data in Database and Hadoop via SQL, R, and Big Data SQL

3. **Embedded R Package Callouts**
   - R Engine(s) spawned by Oracle DB for database-managed parallelism
   - ore.groupApply high performance scoring
   - Efficient data transfer to spawned R engines
   - Emulate map-reduce style algorithms and applications
   - Enables production deployment and automated execution of R scripts
Structured and Unstructured Data Growth

IDC Study: Structured Versus Unstructured Data: The Balance of Power Continues to Shift

“80% of business-relevant information originates in unstructured form, primarily text.”

Structured Versus Unstructured Data: The Balance of Power Continues to Shift

Unstructured Data
Opportunity for Better Insights and Better Actionable Analytics

• Missing from most predictive models
  – Customer comments
  – Emails
  – Customer Service Rep notes
  – Pdfs, Ppts, Word documents, etc.
  – Tweets
  – Physician and Nurse notes
  – Article abstracts
  – Explanations
  – Free form written information that describes more about a situation e.g. a customer’s interest in “discount” and “sale” items, etc. than structured data possibly can. 😊

http://www.zdnet.com/article/unstructured-data-challenge-or-asset/
Oracle Text 12c
Native Capability of every Oracle Database

• Oracle Text uses standard SQL to index, search, and analyze text and documents stored in the Oracle database, in files, and on the web.

• Oracle Text supports multiple languages and uses advanced relevance-ranking technology to improve search quality.

• Oracle Advanced Analytics uses Oracle Text to pre-process ("tokenize") unstructured data for the OAA SQL data mining functions.
Rapidly Build, Evaluate & Deploy Analytical Methodologies
Leveraging a Variety of Data Sources and Types

Consider:
- Demographics
- Past purchases
- Recent purchases
- Comments & tweets

Unstructured data also mined by algorithms

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

Transactional POS data

Modeling Approaches

Generates SQL scripts and workflow API for deployment

Inline predictive model to augment input data
Advanced SQL Engine

Comprehensive SQL Language

Any data: Structured, Semi-structured

Sophisticated query optimization

Joins, aggregations, filters, indexes, query transformations

Smart processing

Parallel, In-memory

Any Answer

Operational, reporting, analytical, predictive

Full-featured Query Processing Engine
Oracle Advanced Analytics—On Premise or Cloud

100% Compatibility Enables Easy Coexistence and Migration

Transparently move workloads and ML/analytical methodologies between On-premise and public cloud
Manage and **Analyze** All Data—SQL & Oracle Big Data SQL

**Structured and Unstructured Data Reservoir**
- JSON data
- HDFS / Hive
- NoSQL
- Spatial and Graph data
- Image and Video data
- Social Media

**Store business-critical data in Oracle**
- Customer data
- Transactional data
- Unstructured documents, comments
- Spatial and Graph data
- Image and Video data
- Social Media

**Data analyzed via SQL / R / GUI**
- R Clients
- SQL Clients
- Oracle Data Miner

**Oracle Big Data SQL + Advanced Analytics**
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.

Engineered Features – Derived attributes/variable that reflect domain knowledge—key to best models

Naïve Guess or Random

Model with 20 variables

Model with 75 variables

Model with 250 variables
Oracle Advanced Analytics

Brief Demos
Oracle Data Miner GUI
Easy to Use for “Citizen Data Scientist”

• Easy to use to define analytical methodologies that can be shared
• SQL Developer Extension
• Workflow API and generates SQL code for immediate deployment
<table>
<thead>
<tr>
<th>CUST_ID</th>
<th>N_TRANS_ATM</th>
<th>CHECKING_AMOUNT</th>
<th>MARRITAL_STATUS</th>
<th>SEX</th>
<th>N_TRANS_CHECK</th>
<th>STATE</th>
<th>HOUSE_OWNERSHIP</th>
<th>MONTHLY_CHECKS</th>
<th>OPENED</th>
<th>SALARY</th>
<th>BANK_FUNDS</th>
<th>SAVINGS</th>
<th>CREDIT_BALANCE</th>
<th>MORTGAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>012464</td>
<td>6</td>
<td>25</td>
<td>MARRIED</td>
<td>F</td>
<td>9</td>
<td>CA</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6,000</td>
<td>63,983</td>
<td>23,022</td>
<td>25,022</td>
<td>546</td>
</tr>
<tr>
<td>012465</td>
<td>0</td>
<td>25</td>
<td>MARRIED</td>
<td>M</td>
<td>2</td>
<td>MI</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0,294.4</td>
<td>59,976</td>
<td>25,000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012466</td>
<td>5</td>
<td>25</td>
<td>WIDOWED</td>
<td>M</td>
<td>2</td>
<td>CA</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>26,477.5</td>
<td>67,994</td>
<td>17,800</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012467</td>
<td>2</td>
<td>7</td>
<td>DIVORCED</td>
<td>M</td>
<td>2</td>
<td>MI</td>
<td>1</td>
<td>1</td>
<td>13,266</td>
<td>71,227</td>
<td>7,600</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012468</td>
<td>4</td>
<td>10</td>
<td>DIVORCED</td>
<td>F</td>
<td>2</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>12,319</td>
<td>70,986</td>
<td>1,300</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012469</td>
<td>4</td>
<td>10,000</td>
<td>DIVORCED</td>
<td>M</td>
<td>2</td>
<td>CA</td>
<td>1</td>
<td>1</td>
<td>1,132,211</td>
<td>58,887</td>
<td>10,000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012471</td>
<td>4</td>
<td>10,943</td>
<td>MARRIED</td>
<td>M</td>
<td>2</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,333</td>
<td>69,335</td>
<td>4,736</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012472</td>
<td>5</td>
<td>20</td>
<td>MARRIED</td>
<td>F</td>
<td>2</td>
<td>FL</td>
<td>1</td>
<td>1</td>
<td>12,200</td>
<td>58,342</td>
<td>2,600</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012473</td>
<td>3</td>
<td>20</td>
<td>MARRIED</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012474</td>
<td>2</td>
<td>20</td>
<td>MARRIED</td>
<td>M</td>
<td>2</td>
<td>CA</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012475</td>
<td>2</td>
<td>20</td>
<td>MARRIED</td>
<td>F</td>
<td>1</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>1,327</td>
<td>62,406</td>
<td>5,405</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012476</td>
<td>2</td>
<td>20</td>
<td>MARRIED</td>
<td>F</td>
<td>2</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012477</td>
<td>2</td>
<td>20</td>
<td>MARRIED</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012478</td>
<td>2</td>
<td>20</td>
<td>MARRIED</td>
<td>F</td>
<td>2</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012479</td>
<td>4</td>
<td>20</td>
<td>MARRIED</td>
<td>M</td>
<td>2</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>3,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>012480</td>
<td>4</td>
<td>20</td>
<td>MARRIED</td>
<td>F</td>
<td>1</td>
<td>NY</td>
<td>1</td>
<td>1</td>
<td>2,185</td>
<td>67,061</td>
<td>1,045</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
<Details algorithm="Naive Bayes" class="Yes">
  
  <Attribute name="LV_Bil" actualValue="MEDIUM" operator="in" range="LOW, MEDIUM, VERY HIGH" weight="0.587" rank="1"/>
  
  <Attribute name="N_TRANS_WFR_BANK" actualValue="2300" operator="greaterThan" value="1419.5" weight="0.963" rank="2"/>
  
  <Attribute name="LV" actualValue="21389.75" operator="greaterThan" value="6861.625" weight="0.848" rank="3"/>
  
  <Attribute name="CAR_OWNERSHIP" actualValue="1" operator="between" range="(0, 0.01]" weight="0.8" rank="5"/>

</Details>
Sharing, Automation and Deployment
Immediately Go to “Productionization” of Analytical Methodologies

- Share ODMr workflows
- Workflow API for 100% automation
  - Immediate deployment of data analyst’s methodologies
- SQL Script Generation
  - Deploy methodology as SQL scripts
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;

Automated Monthly “Application”!

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

Time measure: set timing on;
Predictive analytics EXPLAIN routine

-- Cleanup old output table for repeat runs
BEGIN EXECUTE IMMEDIATE 'DROP TABLE ai_explain_output';
EXCEPTION WHEN OTHERS THEN NULL; END;
/

-- Run the EXPLAIN routine to get attribute importance results
BEGIN
  DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(
    data_table_name => 'mining_data_build_v',
    explain_column_name => 'affinity_card',
    result_table_name => 'ai_explain_output');
END;
/

-- DISPLAY RESULTS

-- List of attribute names ranked by their importance value.
-- The larger the value, the more impact that attribute has
-- on causing variation in the target column.
--
column attribute_name format a40
column explanatory_value format 9.999
SELECT attribute_name, explanatory_value, rank
FROM ai_explain_output
ORDER BY rank, attribute_name;
Given demographic data about a set of customers, predict the customer response to an affinity card program using a classifier based on Decision Trees algorithm.

CREATE A NEW MODEL

-- Build a DT model
BEGIN
    DBMS_DATA_MINING.CREATE_MODEL(
        model_name => 'DT_SH_Clas_sample',
        mining_function => dbms_data_mining.classification,
        data_table_name => 'mining_data_build_v',
        case_id_column_name => 'cust_id',
        target_column_name => 'affinity_card',
        settings_table_name => 'dt_sh_sample_settings');
END;
/

SELECT T.cust_id, S.prediction, S.probability, S.cost
FROM (SELECT cust_id,
    PREDICTION_SET(dt_sh_clas_sample COST MODEL USING *) pset
    FROM mining_data_apply_v
    WHERE cust_id < 100011) T,
     TABLE(T.pset) S
ORDER BY cust_id, S.prediction;
Oracle Advanced Analytics
Real-Time Scoring, Predictions and Recommendations

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

```
Select prediction_probability(FRAUD_1_SVM_1, 'Yes'
    USING 7800 as bank_funds, 125 as checking_amount, 20 as credit_balance, 55 as age, 'Married' as marital_status, 250 as MONEY_MONTLY_OVERDRAWN, 1 as house_ownership)
from dual;
```

Likelihood to respond:

```
PREDICTION_PROB... = 0.8362936507936...
```
Oracle’s Advanced Analytics

Example Customer References
Objectives

- Prevent $200M in losses every year using data to monitor, understand and anticipate fraud

Solution

- We installed OAA analytics for model development during 2014
- When choosing the tools for fraud management, speed is a critical factor
- OAA provided a fast and flexible solution for model building, visualization and integration with production processes

“When choosing the tools for fraud management, speed is a critical factor. Oracle Advance Analytics provided a fast and flexible solution for model building, visualization and integration with production processes.”

– Miguel Barrera, Director of Risk Analytics, Fiserv Inc.
– Julia Minkowski, Risk Analytics Manager, Fiserv Inc.
Data Miner Survey 2016 by Rexer Analytics

While 6 out of 10 data miners report the data is available for analysis within days of capture, the time to deploy the models takes substantially longer. For 60% of the respondents, the deployment time will range between 3 weeks and 1 year.

Ease of Deployment

Everyone forgets about deployment – but is most important component!
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.
UK National Health Service

Combating Healthcare Fraud

Objectives
- Use new insight to help identify cost savings and meet goals
- Identify and prevent healthcare fraud and benefit eligibility errors to save costs
- Leverage existing data to transform business and productivity

Solution
- Identified up to GBP100 million (US$156 million) potentially saved through benefit fraud and error reduction
- Used anomaly detection to uncover fraudulent activity where some dentists split a single course of treatment into multiple parts and presented claims for multiple treatments
- Analyzed billions of records at one time to measure longer-term patient journeys and to analyze drug prescribing patterns to improve patient care

"Oracle Advanced Analytics' data mining capabilities and Oracle Exalytics' performance really impressed us. The overall solution is very fast, and our investment very quickly provided value. We can now do so much more with our data, resulting in significant savings for the NHS as a whole"
– Nina Monckton, Head of Information Services, NHS Business Services Authority

Oracle Exadata Database Machine
Oracle Advanced Analytics
Oracle Exalytics In-Memory Machine
Oracle Endeca Information Discovery
Oracle Business Intelligence EE
DX Marketing
Cloud Based Predictive Analytics/Database Marketing

Objectives

- Cloud-based solution
- Increase revenue
- Reduce time-to-market

Solution

The company considered only two solution vendors --SAS and Oracle to host its consumer data. SAS offered to help build the IT infrastructure from scratch and helped develop a one-year plan. But when they looked at the number of personnel needed to manage the infrastructure including administrators, security specialists and analysts as well as Security & HIPPA compliance needed, Oracle’s DBCS solution looked far more attractive. Hence, they decided to go with Oracle. Oracle’s solution offered:

- Scalability
- Built in analytical tools including data mining.
- Built in HIPPA compliance and security features.
- Required fewer resources --only two analysts --Data Engineer and an expert in Predictive Analytics who now manage the entire eco system.

“Time to market has significantly improved from 4-6 weeks to less than a week with the result the company can bring new clients on board faster. This has helped boost revenues by 25% in the six months since using Oracle’s DBCS.”

– DX Marketing

Oracle Cloud
Oracle Advanced Analytics

DX Marketing Expands Customer Acquisition with Oracle Cloud

– YouTube video
Zagrebačka Bank (biggest bank in Croatia)

**Increases Cash Loans by 15% Within 18 Months of Deployment**

### Objectives
- Needed to speed up entire advanced analytics process; data prep was taking 3 days; model building 24 hours
- Faster time to “actionable analytics” for Credit Risk Modeling and Targeted Customer Campaigns

### Solution
- Zaba migrated from SAS to the Oracle Advanced Analytics platform for statistical modeling and predictive analytics
- Increased prediction performance by leveraging the security, reliability, performance, and scalability of Oracle Database and Oracle Advanced Analytics for predictive analytics—running data preparation, transformation, model building, and model scoring within the database

> “With Oracle Advanced Analytics we execute computations on thousands of attributes in parallel—impossible with open-source R. Analyzing in Oracle Database without moving data increases our agility. Oracle Advanced Analytics enables us to make quality decisions on time, increasing our cash loans business 15%.”

-- Jadranka Novoselovic, Head of BI Dev., Zagrebačka Bank

> “We chose Oracle because our entire data modeling process runs on the same machine with the highest performance and level of integration. With Oracle Database we simply switched on the Oracle Advanced Analytics option and needed no new tools,”

-- Sinisa Behin, ICT coordinator at BI Dev. Zagrebačka Bank

ZabaBank Oracle Customer Snapshot on OTN
Objectives

- Store development dashboards to identify opportunities
- 8 M daily transactions, ~25M transaction detail lines
- 20 TB data warehouse size, sales data about 10 TB
- Market basket analysis and customer loyalty & segmentation

Solution

- Exadata Engineered System
- Oracle Advanced Analytics Option
- Market Basket Analysis, Clustering, Classification, Segmentation, Loyalty Analysis

“Exponential growth in combinations with each hierarchy. 2 years of pre-computed Market Baskets and associated sales measures for reporting. Nightly compute within ETL window data with 1 day latency.”

– Dunkin Brands, Mahesh Jagannath, Senior Manager, Business Intelligence

(Excerpts from Dunkin Brands presentation at Oracle Open World 2014)
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

Improve Gaming Experience with Big Data Analytics

- Manage and analyze up to 300 billion events per day
- Understand and segment players
- Quickly correct game play problems
- Deployed Oracle Advanced Analytics and Oracle R Advanced Analytics for Hadoop on Oracle Big Data Appliance and Oracle Database Appliance

62%

Increase in revenue in one region by tailoring messages and playing experience
An Post

Boosts Retail & Postal Services with Big Data & Analytics Platform

Objectives

- Provide a scalable big data and analytics platform to manage millions of daily transactions, facilitate new postal services, and align with market needs and the growth in post office retail services

Solution

- Deployed a big data and analytics platform using Oracle Exadata Database Machine, Oracle Advanced Analytics, ... to ..., combat fraud, and enable readily-available enterprise wide business ... within core areas such as mails, parcels, and retail
- Enabled An Post to rapidly analyze and respond ... using Oracle Advanced Analytics...
- Vertice (partner) provided expert services including ... in-depth product and data mining expertise services

“With our Oracle solution for Big Data and Analytics we now have an analytics platform that has completely transformed our service delivery model. It provides An Post with a single source of truth while automatically consolidating data from all post offices nationwide.”

– John Cronin, Group Chief Information Officer, An Post Limited

“Big data is a vital part of our future,” says John Cronin, CIO of Ireland’s postal network.

https://www.siliconrepublic.com/enterprise/2016/02/05/an-post-oracle-john-cronin-five-minute-cio
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
R: Transparency through function overloading

Invoke in-database aggregation function

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

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

```r
head(aggdata)
  Group.1 x
 Group.1 x
1 ABE  237
2 ABI  34
3 ABQ 1357
4 ABY  10
5 ACK  3
6 ACT  33
```
R: Transparency through function overloading
Invoke in-database Data Mining model (Support Vector Machine)

R Console

Oracle Distribution of R version 3.0.1 (2012-06-22) -- "Good Sport"

> svm_mod <- ore.odmSVM(formula = BUY ~ INCOME + YRS_CUST + MARITAL_STATUS, data=CUST, type = "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:

<table>
<thead>
<tr>
<th>class</th>
<th>variable</th>
<th>value</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>INCOME</td>
<td>5.204561e-05</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>MARITAL_STATUS</td>
<td>M-4.531359e-05</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>MARITAL_STATUS</td>
<td>S-4.531359e-05</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>YRS_CUST</td>
<td>1.264948e-04</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>(Intercept)</td>
<td>9.999269e-01</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INCOME</td>
<td>2.032340e-05</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MARITAL_STATUS</td>
<td>M-2.636552e-06</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MARITAL_STATUS</td>
<td>S-2.636552e-06</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>YRS_CUST</td>
<td>-1.588211e-04</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Intercept)</td>
<td>-9.999324e-01</td>
<td></td>
</tr>
</tbody>
</table>
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
Using Hadoop and HIVE Integration, plus R Engine and Open-Source R Packages

Hadoop Cluster
with Oracle R Advanced Analytics for Hadoop (ORAAH)

- R interface to HQL Basic Statistics, HQL Data Prep, Joins and View creation
- Parallel, distributed algorithms:
  - MLP Neural Nets*, GLM*, LM, PCA
  - k-Means, NMF, LMF
  - * Spark-Caching enabled
- Use of Open-source R packages via custom R Mappers / Reducers

R Client
Oracle R Advanced Analytics for Hadoop

R Analytics

Oracle Database
with Advanced Analytics option

SQL Client
SQL Developer
Other SQL Apps

Copyright © 2016 Oracle and/or its affiliates. All rights reserved.
# Oracle R Advanced Analytics for Hadoop

## AA Algorithms in a Hadoop Cluster: Map-Reduce and Spark (2.6)

### Classification
- GLM ORAAH
- Logistic Regression ORAAH
- Random Forests
- Decision Trees
- Support Vector Machines

### Regression
- MLP Neural Networks ORAAH
- LASSO
- Ridge Regression
- Support Vector Machines
- Random Forest
- Linear Regression

### Clustering
- Hierarchical k-Means
- Hierarchical k-Means

### Feature Extraction
- Non-negative Matrix Factorization
- Collaborative Filtering (LMF)

### Basic Statistics
- Correlation/Covariance

### Attribute Importance
- Principal Components Analysis
- Principal Components Analysis

---

*Oracle and/or its affiliates. All rights reserved.*
Invoke ORAAH custom parallel distributed GLM Model using Spark Caching

ORAAH: Machine Learning in Spark against HDFS data

> spark.connect("yarn-client", memory="24g")

# Attaches the HDFS file for use within R
> ont1bi <- hdfs.attach("/user/oracle/ontime_1bi")

# Formula definition: Cancelled flights (0 or 1) based on other attributes
> form_oraah_glm2 <- CANCELLED ~ DISTANCE + ORIGIN + DEST + F(YEAR) + F(MONTH) +
  + F(DAYOFMONTH) + F(DAYOFWEEK)

> system.time(m_spark_glm <- orch glm2(formula=form_oraah_glm2, ont1bi))

ORCH GLM: processed 6 factor variables, 25.806 sec
ORCH GLM: created model matrix, 100128 partitions, 32.871 sec
ORCH GLM: iter 1, deviance 1.38433414089348300E+09, elapsed time 9.582 sec
ORCH GLM: iter 2, deviance 3.39315388583931150E+08, elapsed time 9.213 sec
ORCH GLM: iter 3, deviance 2.06855738812683250E+08, elapsed time 9.218 sec
ORCH GLM: iter 4, deviance 1.75868100359263200E+08, elapsed time 9.104 sec
ORCH GLM: iter 5, deviance 1.70023181759611580E+08, elapsed time 9.132 sec
ORCH GLM: iter 6, deviance 1.69476890425481350E+08, elapsed time 9.124 sec
ORCH GLM: iter 7, deviance 1.69467586045954760E+08, elapsed time 9.077 sec
ORCH GLM: iter 8, deviance 1.69467574351380850E+08, elapsed time 9.164 sec

user system elapsed
84.107  5.606 143.591
Oracle’s Advanced Analytics
Predictive Applications + OBIEE Integration
Enabling “Predictive” Enterprise Applications

Oracle Applications Using Oracle Advanced Analytics—Partial List

• **Oracle HCM Fusion**
  – Employee turnover and performance prediction and “What if?” analysis

• **Oracle CRM Fusion**
  – Prediction of sales opportunities, what to sell, amount, timing, etc.

• **Oracle Industry Data Models**
  – **Communications Data Model** churn prediction, segmentation, profiling, etc.
  – **Retail Data Model** loyalty and market basket analysis
  – **Airline Data Model** analysis frequent flyers, loyalty, etc.
  – **Utilities Data Model** customer churn, cross-sell, loyalty, etc.

• **Oracle Retail Customer Analytics**
  – “Shopping cart analysis” and next best offers

• **Oracle Customer Support**
  – Predictive Incident Monitoring (PIM)

• **Oracle Spend Classification**
  – Real-time and batch flagging of noncompliance and anomalies in expense submissions

• **Oracle FinServ Analytic Applications**
  – Customer Insight, Enterprise Risk Management, Enterprise Performance, Financial Crime and Compliance

• **Oracle Adaptive Access Manager**
  – Real-time security and fraud analytics
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

Oracle Communications Industry Data Model
Example Predictive Analytics Application

Link to OCDM on OTN
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

Link to OCDM on OTN
OCDM Telco Churn Enhanced by SNA Analysis

- 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

Link to OCDM SNA Documentation
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


Link to Oracle HCM on O.com
HCM Predictive Workforce demo
**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

Link to Oracle HCM on O.com

HCM Predictive Workforce demo

Copyright © 2016 Oracle and/or its affiliates. All rights reserved.
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

Link to Oracle HCM on O.com
HCM Predictive Workforce demo
Sales Predictor helps sales reps answer critical sales questions:

- Which products should be offered to a customer?
- Who are the customers buying products?
- What are the reasons a product is being bought?

Sales Predictor offers product recommendations that have a higher likelihood of being converted to a win.

Link to Oracle CRM SPE on O.com
Oracle Retail Market Basket Insights Cloud Service
Market Basket Analysis

Pre-Built Market Basket Analysis

- Gain actionable insight into your shoppers' behavior.
- Pre-built market-basket analysis identifies product affinities

[Diagram of pre-built market basket analysis process]

Link to Oracle Retail MBA on O.com

Copyright © 2016 Oracle and/or its affiliates. All rights reserved.
Pre-Built Customer Clustering Models

• Gain actionable insight into your customer’s behavior.
• Pre-built clustering models identify hidden customer segments.
Oracle Spend Classification
Automated Pre-built Classification models for Spend Assignment

Oracle Procurement and Spend Analytics Module

• Uses Oracle Advanced Analytics’ to automatically categorize spend data
• Uses OBIA invoice, purchasing, and requisition data
• Oracle Spend Classification predicts purchasing categories for spend lines, it reassigns categories in P&SA where it has a high confidence rating of a successful match.

Link to Oracle Spend Class on O.com
Oracle Retail Industry Data Model
Predictive Analytics Application

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
Getting started
Oracle Advanced Analytics Overview:
- OAA presentation — Big Data Analytics with Oracle Advanced Analytics — or just watch YouTube video presentation and demo(s)
- Big Data Analytics with Oracle Advanced Analytics: Making Big Data and Analytics Simple white paper on OTN
- Oracle Internal OAA Product Management Wiki and Workspace
- Oracle Advanced Analytics Customer Successes

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 OAA/Oracle Data Miner Workflow GUI Online (free) Tutorial Series on OTN
- Link to OAA/Oracle R Enterprise (free) Tutorial Series on OTN
- Link to Free Oracle Advanced Analytics "Test Drives" on Oracle Cloud via Vlamis Partner
- Link to Getting Started w/ ODM blog entry
- Link to New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course.
- Oracle Data Mining Sample Code Examples

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
- Oracle R Advanced Analytics for Hadoop (ORAAH) on OTN
- Business Intelligence, Warehousing & Analytics — BIWA Summit’17, Jan 31, Feb 1 & 2, 2017 at Oracle HQ Conference Center (w/ links to customer presentations)
Take a Test Drive on the Oracle Public Cloud!
Vlamis Software, Oracle Partner Offers FREE Test Drives on the Oracle Cloud

- **Step 1**—Fill out request
  - Go to http://www.vlamis.com/td/

- **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

Try Business Intelligence and Advanced Analytics on the Oracle Cloud for Free!

We've partnered with Oracle to provide you a shortcut to see what the end user experience of the Oracle Cloud is like. The Oracle Cloud with custom processes that we have created give us a flexible and resilient demonstration environment that we can let you try for 24 hours, for free, by filling out this form.
Books on Oracle Advanced Analytics

Book available on Amazon
Predictive Analytics Using Oracle Data Miner: Develop for ODM in SQL & PL/SQL

Book available on Amazon
Using R to Unlock the Value of Big Data
BIWA SUMMIT 2017

The Oracle Big Data + Analytics User Conference

January 31, Feb 1 & 2, 2017

Including Oracle Spatial Summit

---

Mark your Calendars now!

Call for Abstracts coming soon...

Oracle Conference Center at Oracle HQ Campus, Redwood Shores, CA

- Hands-on-Labs
- Customer stories
- Educational sessions
- Oracle Keynote
- Presentations covering: Advanced Analytics, Big Data, Business Intelligence, Cloud, Data Warehousing and Integration, Spatial and Graph, SQL
- Networking with product management and development professionals
Oracle Advanced Analytics 12.2, Oracle Data Miner 4.2 and ORAAH 2.6

New Features
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.
Oracle Advanced Analytics 12.2

New Oracle Database Features

• **Major Performance Improvements for Algorithms**
  – New parallel model build / apply infrastructure to enable faster algorithm introduction
  – Scale to larger data volumes found in big data and cloud use cases

• **Unsupervised Feature Selection**
  – Uses unsupervised pair-wise Kullback-Leibler divergence (correlations analysis - numeric and categorical attributes) to find highest “information containing” attributes

• **Association Rules Enhancements**
  – Adds calculation of values associated with AR rules such as sales amount to indicate the value of co-occurring items in baskets

• **Partitioned Models**
  – Instead of building, naming and referencing 10s or 1000s of models, a partitioned model will organize and represent these multiple models as partitions in a single model entity
• **Explicit Semantic Analysis (ESA) algorithm**
  – Useful technique for extracting meaningful, interpretable features, better than LDA
  – Use ESA for document similarity and topic identification
  – Wikipedia provides a large corpus of existing documents to provide sensible features and topics
  – Document 1
  – ‘Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa’ida, including bin Laden. The money enabled al-Qa’ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia.'

Document 2
– ‘The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed “and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court,” the official Saudi Press Agency quoted an unidentified ministry official as saying.'

**ESA Similarity 0.62**
Oracle Advanced Analytics 12.2
New Oracle Database Features

• Extensibility for R Models
  – Register R models as in-database models for build, apply, settings, and viewing
  – Supports data with “nested” attributes, handling text and aggregated transactional data
  – *Extends ease of advanced analytics development from R to Oracle Database*
  – *Enables R users to roll out new analytics and more rapidly take advantage of existing R packages*
Oracle’s Advanced Analytics 12.2
In-Database Data Mining Algorithms*—SQL & R & GUI Access

<table>
<thead>
<tr>
<th>Classification</th>
<th>Clustering</th>
<th>Predictive Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>Hierarchical k-Means</td>
<td>Clustering</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Orthogonal Partitioning Clustering</td>
<td>Regression</td>
</tr>
<tr>
<td>(GLM)</td>
<td>Expectation-Maximization</td>
<td>Anomaly Detection</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td></td>
<td>Feature Extraction &amp; Creation</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td></td>
<td>Market Basket Analysis</td>
</tr>
<tr>
<td>Multiple Regression</td>
<td>Minimum Description Length</td>
<td></td>
</tr>
<tr>
<td>(GLM)</td>
<td>Unsupervised pair-wise KL div.</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalized Linear Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-Layer Neural Networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stepwise Linear Regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attribute Importance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum Description Length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unsupervised pair-wise KL div.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anomaly Detection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 Class Support Vector Machine</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time Series</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Single &amp; Double Exp. Smoothing</td>
<td></td>
</tr>
</tbody>
</table>

* Supports Partitioned models, Text Mining and Nested Data
### Oracle Advanced Analytics 12.2

Preliminary Model Build Times Performances #s

<table>
<thead>
<tr>
<th>OAA 12.2 Algorithm</th>
<th>Rows (Ms)</th>
<th>T7-4 Model Build Time (Secs / Degree of Parallelism)</th>
<th>X5-4 Model Build Time (Secs / Degree of Parallelism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributes Importance</td>
<td>640</td>
<td>28s / 512</td>
<td>44s / 72</td>
</tr>
<tr>
<td>Attribute Importance</td>
<td>320</td>
<td>16s / 256</td>
<td>23s / 72</td>
</tr>
<tr>
<td>Attributes Importance</td>
<td>159</td>
<td>10s / 256</td>
<td>13s / 72</td>
</tr>
<tr>
<td>K Means</td>
<td>640</td>
<td>161s / 256</td>
<td>268s / 144</td>
</tr>
<tr>
<td>K Means</td>
<td>320</td>
<td>87s / 256</td>
<td>132s / 144</td>
</tr>
<tr>
<td>K Means</td>
<td>159</td>
<td>51s / 256</td>
<td>62s / 144</td>
</tr>
<tr>
<td>Expectation Maximization</td>
<td>159</td>
<td>455s / 512</td>
<td>588s / 144</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>320</td>
<td>17s / 256</td>
<td>23s / 72</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>159</td>
<td>10s / 256</td>
<td>13s / 72</td>
</tr>
<tr>
<td>GLM Classification</td>
<td>640</td>
<td>154s / 512</td>
<td>363s / 144</td>
</tr>
<tr>
<td>GLM Classification</td>
<td>320</td>
<td>92s / 512</td>
<td>260s / 144</td>
</tr>
<tr>
<td>GLM Classification</td>
<td>159</td>
<td>54s / 512</td>
<td>138s / 144</td>
</tr>
</tbody>
</table>

The way to read their results is that they compare 2 chips: X5 (Intel and Linux) and T7 (Sparc and Solaris). They are measuring scalability (time in seconds) with increase degree of parallelism (dop). The data also has high cardinality categorical columns which translates in 9K mining attributes (when algorithms require explosion). There are no comparisons to 12.1 and it is fair to say that the 12.1 algorithms could not run on data of this size.
# Oracle Advanced Analytics 12.2

## Preliminary Model Build Times Performances #s

<table>
<thead>
<tr>
<th>OAA 12.2 Algorithm</th>
<th>Rows (Ms)</th>
<th>T7-4 Model Build Time (Secs / Degree of Parallelism)</th>
<th>X5-4 Model Build Time (Secs / Degree of Parallelism)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM Regression</td>
<td>640</td>
<td>55s / 512</td>
<td>93s / 144</td>
</tr>
<tr>
<td>GLM Regression</td>
<td>320</td>
<td>35s / 256</td>
<td>50s / 144</td>
</tr>
<tr>
<td>GLM Regression</td>
<td>159</td>
<td>18s / 256</td>
<td>29s / 72</td>
</tr>
<tr>
<td>SVM (IPM solver)</td>
<td>640</td>
<td>404s / 512</td>
<td>1411s / 144</td>
</tr>
<tr>
<td>SVM (IPM solver)</td>
<td>320</td>
<td>385s / 512</td>
<td>1263s / 72</td>
</tr>
<tr>
<td>SVM (IPM solver)</td>
<td>159</td>
<td>220s / 512</td>
<td>608s / 72</td>
</tr>
<tr>
<td>SVM (SGD solver)</td>
<td>640</td>
<td>84s / 256</td>
<td>188s / 72</td>
</tr>
<tr>
<td>SVM (SGD solver)</td>
<td>320</td>
<td>55s / 256</td>
<td>116s / 72</td>
</tr>
<tr>
<td>SVM (SGD solver)</td>
<td>159</td>
<td>36s / 128</td>
<td>71s / 36</td>
</tr>
</tbody>
</table>

The way to read their results is that they compare 2 chips: X5 (Intel and Linux) and T7 (Sparc and Solaris). They are measuring scalability (time in seconds) with increase degree of parallelism (dop). The data also has high cardinality categorical columns which translates in 9K mining attributes (when algorithms require explosion). There are no comparisons to 12.1 and it is fair to say that the 12.1 algorithms could not run on data of this size.
Oracle Data Miner

4.2 New Features
Oracle Data Miner 4.2

New Features for OAA

• Add/Expose all 12.2 features in Oracle Data Miner UI
Oracle Data Miner 4.2

New Features for OAA

• Add/Expose all 12.2 features in Oracle Data Miner UI
  – Partitioned Models
  – Association Rules filter items (before & after model build) & aggregation Columns
  – Unsupervised Feature Selection & unsupervised pair-wise dependencies/correlations
  – Explicit Feature Extraction node & viewer (tag cloud like output) w/ sample Wiki table
  – Feature Compare Node - compare two data flows for similarity using FE node (explicit & latent). Data flow can be a single record entered manually
  – R Build Node - Build, Score and Model Detail functions are defined by selecting an existing R script registered in DB
  – In-Memory performance options with existing Parallel options
Oracle Data Miner 4.2

New Features for OAA

- Miscellaneous Improvements based on customer use cases/feedback
  - Workflow Scheduler
    - Wide selection of scheduling options available
    - Integration with Email Notification for a workflow start/stop/failure events
    - Users can now view workflow details completely when a workflow is running or scheduled to run (Access was restricted in prior releases)
  - Aggregation Node supports Date, TimeStamp aggregations
  - Tables, Views, User, Column Names, etc. changed from 30 to 128 characters (configurable to remain at 30)
  - Optimized workflow polling frequency
  - Greater DBA control of workflow execution and behavior
  - Accept RAW, ROWID, UROWID, URITYPE data types
Previewing a 4.2 Feature

Workflow Scheduler
### Classification
- GLM ORAAH
- Logistic Regression ORAAH
- Random Forests
- Decision Trees
- Support Vector Machines

### Regression
- MLP Neural Networks ORAAH
- LASSO
- Ridge Regression
- Support Vector Machines
- Random Forest
- Linear Regression

### Basic Statistics
- Correlation/Covariance

### Feature Extraction
- Non-negative Matrix Factorization
- Collaborative Filtering (LMF)

### Clustering
- Hierarchical k-Means

### Attribute Importance
- Principal Components Analysis

Oracle R Advanced Analytics for Hadoop

AA Algorithms in a Hadoop Cluster: Map-Reduce and Spark (2.6)
Roadmap Architecture

GUI  SQL  R

Algorithms
Common core, parallel, distributed

Hadoop  Spark  Relational  Cloud