Big Data Analytics with Oracle Advanced Analytics In-Database Option

Charlie Berger
Sr. Director Product Management, Data Mining and Advanced Analytics
charlie.berger@oracle.com
www.twitter.com/CharlieDataMine
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
"Big Data" → "Big Data Analytics"

1.8 trillion gigabytes of data was created in 2011...

- More than 90% is unstructured data
- Approx. 500 quadrillion files
- Quantity doubles every 2 years

"There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing."

- Google CEO Eric Schmidt

Requires capability to rapidly:

✅ Collect and integrate data
✅ Understand data & their relationships
✅ Respond and take action

Source: IDC 2011
Oracle Big Data Platform

Oracle Big Data Appliance
Optimized for Hadoop, R, and NoSQL Processing

Oracle Big Data Connectors

Oracle Exadata
“System of Record” Optimized for DW/OLTP

Oracle Exalytics
Optimized for Analytics & In-Memory Workloads

Stream | Acquire | Organize | Discover & Analyze

Hadoop
Open Source R
Oracle NoSQL Database
Applications

Oracle Big Data Connectors

Oracle Advanced Analytics
Data Warehouse
Oracle Database

In-Database Analytics

Oracle Enterprise Performance Management
Oracle Business Intelligence Applications
Oracle Business Intelligence Tools
Oracle Endeca Information Discovery
“Without proper analysis, it's just data; ...not useful actionable information ...something that you can exploit today ...something that your competitor may not have yet discovered.”

**Charlie Berger**
Sr. Director, Product Management, Oracle Data Mining and Advanced Analytics
Oracle Corporation
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
Data Mining Provides
Better Information, Valuable Insights and Predictions

Source: Inspired from Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff
Finding Needles in Haystacks

- Haystacks are usually BIG
- Needles are typically small and rare
Challenge: Finding Anomalies

• Look for what is “different”
• Single observed value, taken alone, may seem “normal”
• Consider multiple attributes simultaneously
• Taken collectively, a record may appear to be anomalous
Data Mining & Predictive Analytics

Example Use Cases for Advanced Analytics

- Targeting the right customer with the right offer
- Discovering hidden customer segments
- Finding most profitable selling opportunities
- Anticipating and preventing customer churn
- Exploiting the full 360 degree customer opportunity
- Security and suspicious activity detection
- Understanding sentiments in customer conversations
- Reducing medical errors & improving quality of health
- Understanding influencers in social networks
Oracle Advanced Analytics Option
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

- **Better Decisions with Deeper Insights & Predictive Analytics**
  - Understand and predict customer behavior for churn, fraud, cross-sell, etc. problems

- **Easy to Use**
  - Data analysts: Oracle Data Miner work flow GUI (part of SQL Developer)
  - Data scientists: SQL and R languages supported
  - DBA: Oracle Database EE Option

- **Comprehensive Analytics on a Simple Architecture**
  - Performance and scalability of the Oracle Database
  - Lowest Total Costs of Ownership; no need for separate analytical servers
Why Oracle Advanced Analytics?

Differentiating Features

✔ Fastest Way to Deliver Enterprise Predictive Analytics Applications
  ▪ Integrated with OBIEE and any application that uses SQL queries

✔ Performance and Scalability
  ▪ Leverages power and scalability of Oracle Database.

✔ Lowest Total Costs of Ownership
  ▪ No need for separate analytical servers
Oracle Advanced Analytics Value Proposition

Value Proposition
- Fastest path from data to insights
  - Fastest analytical development
  - Fastest in-database scoring engine on the planet
- Flexible deployment options for analytics
- Lowest TCO by eliminating data duplication
- Secure, Scalable and Manageable

Data remains in the Database

Data preparation for analytics is automated

Scalable distributed-parallel implementation of machine learning techniques in the database

Scalable implementation of R programming language in-database

Flexible interface options – SQL, R, IDE, GUI

Fastest and most Flexible analytic deployment options

Can import 3rd party models
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 Data Miner 11g Release 2 GUI

Anomaly Detection—Simple Conceptual Workflow

Train on “normal” records
Apply model and sort on likelihood to be “different”
Fraud Prediction Demo

drop table CLAIMS_SET;
exec dbms_data_mining.drop_model('CLAIMSMODEL');
create table CLAIMS_SET (setting_name varchar2(30), setting_value varchar2(4000));
insert into CLAIMS_SET values ('ALGO_NAME','ALGO_SUPPORT VECTOR MACHINES');
insert into CLAIMS_SET values ('PREP_AUTO','ON');
commit;
begin
  dbms_data_mining.create_model('CLAIMSMODEL', 'CLASSIFICATION', 'CLAIMS', 'POLICYNUMBER', null, 'CLAIMS_SET');
end;
/

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

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

Automated Monthly “Application”! Just add:
Create
View CLAIMS2_30
As
Select * from CLAIMS2
Where mydate > SYSDATE – 30
Better Information for OBI EE Reports and Dashboards

### Most Suspicious Claims

<table>
<thead>
<tr>
<th>POLICYNUMBER</th>
<th>PREDICTION</th>
<th>PROBABILITY</th>
<th>DEDUCTIBLE</th>
<th>DEPENDIBLE * PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>13465.00</td>
<td>0.00</td>
<td>0.67</td>
<td>500.00</td>
<td>336.40</td>
</tr>
<tr>
<td>14485.00</td>
<td>0.00</td>
<td>0.65</td>
<td>400.00</td>
<td>260.13</td>
</tr>
<tr>
<td>6532.00</td>
<td>0.00</td>
<td>0.64</td>
<td>400.00</td>
<td>257.69</td>
</tr>
<tr>
<td>12631.00</td>
<td>0.00</td>
<td>0.63</td>
<td>400.00</td>
<td>251.91</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>0.63</td>
<td>300.00</td>
<td>188.81</td>
</tr>
<tr>
<td>15629.00</td>
<td>0.00</td>
<td>0.63</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>11015.00</td>
<td>0.00</td>
<td>0.63</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>11822.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>4568.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>12650.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>2249.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>12307.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>3331.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>3863.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>13677.00</td>
<td>0.00</td>
<td>0.62</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>11102.00</td>
<td>0.00</td>
<td>0.62</td>
<td>700.00</td>
<td></td>
</tr>
<tr>
<td>13294.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>5284.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>2291.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>3440.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>5139.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>11060.00</td>
<td>0.00</td>
<td>0.61</td>
<td>700.00</td>
<td></td>
</tr>
<tr>
<td>11119.00</td>
<td>0.00</td>
<td>0.61</td>
<td>500.00</td>
<td></td>
</tr>
<tr>
<td>11677.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>1245.00</td>
<td>0.00</td>
<td>0.61</td>
<td>400.00</td>
<td></td>
</tr>
</tbody>
</table>

### Number Of Predictions

<table>
<thead>
<tr>
<th>PREDICTION</th>
<th>COUNT(PREDICTION)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH</td>
<td>5767</td>
</tr>
<tr>
<td>LOW</td>
<td>1337</td>
</tr>
<tr>
<td>MEDIUM</td>
<td>4932</td>
</tr>
<tr>
<td>VERY HIGH</td>
<td>1324</td>
</tr>
</tbody>
</table>

OAA’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools.
Financial Sector/Accounting/Expenses

Anomaly Detection

Simple Fraud Detection Methodology—1-Class SVM

More Sophisticated Fraud Detection Methodology—Clustering + 1-Class SVM
Oracle Advanced Analytics

More Details

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

```sql
Select prediction_probability(CLAS_DT_1_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_MONLY_OVERFLOW, 1 as house_ownership)
from dual;
```

Likelihood to respond:

<table>
<thead>
<tr>
<th>Query Result</th>
<th>All Rows Fetched: 1 in 0 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDICTION_PROB...</td>
<td>0.8362936507936...</td>
</tr>
</tbody>
</table>
Enabling Predictive Applications

Example Applications Using Oracle Advanced Analytics

- **Human Capital Management**
  - **Predictive Workforce**—employee turnover and performance prediction and “What if?” analysis

- **CRM**
  - **Sales Prediction Engine**—prediction of sales opportunities, what to sell, amount, timing, etc.

- **Supply Chain Management**
  - **Spend Classification**—real-time flagging of noncompliance and anomalies in expense submission

- **Identity Management**
  - **Oracle Adaptive Access Manager**—real-time security and fraud analytics

- **Retail Analytics**
  - **Oracle Retail Customer Analytics**—”shopping cart analysis” and next best offers

- **Customer Support**
  - **Predictive Incident Monitoring (PIM)** Customer Service offering for Database customers

- **Manufacturing**
  - Response surface modeling in chip design

- **Predictive capabilities in Oracle Industry Data Models**
  - **Communications Data Model** implements churn prediction, segmentation, profiling, etc.
  - **Retail Data Model** implements loyalty and market basket analysis
  - **Airline Data Model** implements analysis frequent flyers, loyalty, etc.
Oracle Communications Industry Data Model
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

OAA’s clustering and predictions available in-DB for OBIEE
Integrated Business Intelligence
Integrate a range of in-DB SQL & R Predictive Analytics & Graphics

- In-database construction of predictive models that predict customer behavior
- OBIEE’s integrated spatial mapping shows where

Customer “most likely” to be HIGH and VERY HIGH value customer in the future
Integration with Oracle BI EE

Oracle BI EE defines results for end user presentation

Oracle Data Mining results available to Oracle BI EE administrators
• Segment customers into profiles or clusters to more proactively manager “best customers”, like churners, likely fraudsters, etc.

• Drill through for detail to individual records, predictions and profiles
Healthcare Example

Top At-Risk Factors, Guided Drill-Through for Detail

• Given patient hospital admissions and claims history for several years, predict which patients are at highest risk of dying.
• Using OBI EE, select OAA model insights and predictions and define interactive Dashboards with optional drill-through for detail.
Fusion HCM Predictive Analytics
Built-in Predictive Analytics

Oracle Advanced Analytics factory-installed predictive analytics show employees likely to leave, top reasons, expected performance and real-time "What if?" analysis.
Factors associated with Employee’s predicted departure
Oracle Data Miner GUI
SQL Developer 3.2 Extension—Free OTN Download

- **Easy to Use**
  - Oracle Data Miner GUI for data analysts
  - Explore data—discover new insights
  - “Work flow” paradigm for analytical methodologies

- **Powerful**
  - Multiple algorithms & data transformations
  - Runs 100% in-DB
  - Build, evaluate and apply data mining models

- **Automate and Deploy**
  - Generate and deploy SQL scripts for automation
  - Share analytical workflows
Oracle Data Miner GUI
Oracle Data Miner Nodes — Partial List

Tables and Views

Transformations

Explore Data

Modeling

Text
Insurance
Identify “Likely Insurance Buyers” and their Profiles

OAA work flows capture analytical process and generates SQL code for deployment
Oracle Advanced Analytics

Data Mining Unstructured Data

- Mines unstructured i.e. “text” data
- Include text and comments in models
- Cluster and classify documents
- Oracle Text used to preprocess unstructured text
SQL predicates and OAA models are pushed to storage level for execution.

For example, find the US customers likely to churn:

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

## SQL Data Mining Algorithms

<table>
<thead>
<tr>
<th><strong>Problem</strong></th>
<th><strong>Algorithms</strong></th>
<th><strong>Applicability</strong></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 Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>Multiple Regression (GLM)</td>
<td>Classical statistical technique</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
<td>Wide / narrow data / text</td>
</tr>
<tr>
<td><strong>Anomaly</strong></td>
<td>One Class SVM</td>
<td>Lack examples of target field</td>
</tr>
<tr>
<td><strong>Detection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Attribute</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Importance</strong></td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td><strong>Association</strong></td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td><strong>Rules</strong></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td></td>
<td>Apriori</td>
<td></td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>Hierarchical K-Means</td>
<td>Product grouping</td>
</tr>
<tr>
<td></td>
<td>Hierarchical O-Cluster</td>
<td>Text mining</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gene and protein analysis</td>
</tr>
<tr>
<td><strong>Feature</strong></td>
<td>Nonnegative Matrix Factorization</td>
<td>Text analysis</td>
</tr>
<tr>
<td><strong>Extraction</strong></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>

## Oracle Advanced Analytics

- **SQL Data Mining Algorithms**
- **Logistic Regression (GLM)**
- **Decision Trees**
- **Naïve Bayes**
- **Support Vector Machine**
- **Multiple Regression (GLM)**
- **Support Vector Machine**
- **One Class SVM**
- **Minimum Description Length (MDL)**
- **Apriori**
- **Hierarchical K-Means**
- **Hierarchical O-Cluster**
- **Nonnegative Matrix Factorization**

## Applicability

- **Classical statistical technique**
- **Popular / Rules / transparency**
- **Embedded app**
- **Wide / narrow data / text**
- **Market basket analysis**
- **Link analysis**
- **Product grouping**
- **Text mining**
- **Gene and protein analysis**
- **Text analysis**
- **Feature reduction**
Oracle Advanced Analytics
SQL Statistics and SQL Analytics (free)

- **Ranking functions**
  - rank, dense_rank, cume_dist, percent_rank, ntile
- **Window Aggregate functions**
  (moving & cumulative)
  - Avg, sum, min, max, count, variance, stddev, first_value, last_value
- **LAG/LEAD functions**
  - Direct inter-row reference using offsets
- **Reporting Aggregate functions**
  - Sum, avg, min, max, variance, stddev, count, ratio_to_report
- **Statistical Aggregates**
  - Correlation, linear regression family, covariance
- **Linear regression**
  - Fitting of an ordinary-least-squares regression line to a set of number pairs.
  - Frequently combined with the COVAR_POP, COVAR_SAMP, and CORR functions

- **Descriptive Statistics**
  - DBMS_STAT_FUNCS: summarizes numerical columns of a table and returns count, min, max, range, mean, median, stats_mode, variance, standard deviation, quantile values, +/- n sigma values, top/bottom 5 values
- **Correlations**
  - Pearson's correlation coefficients, Spearman's and Kendall's (both nonparametric).
- **Cross Tabs**
  - Enhanced with % statistics: chi squared, phi coefficient, Cramer's V, contingency coefficient, Cohen's kappa
- **Hypothesis Testing**
  - Student t-test, F-test, Binomial test, Wilcoxon Signed Ranks test, Chi-square, Mann Whitney test, Kolmogorov-Smirnov test, One-way ANOVA
- **Distribution Fitting**
  - Kolmogorov-Smirnov Test, Anderson-Darling Test, Chi-Squared Test, Normal, Uniform, Weibull, Exponential

*Note: Statistics and SQL Analytics are included in Oracle Database Standard Edition and Enterprise Edition*
Independent Samples T-Test
(Pooled Variances)

- Query compares the mean of AMOUNT_SOLD between MEN and WOMEN within CUST_INCOME_LEVEL ranges. Returns observed t value and its related two-sided significance

```
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;
```
Oracle’s R Strategic Offerings
Deliver enterprise-level advanced analytics based on R environment

- **Oracle R Distribution**
  - Free download, pre-installed on Oracle Big Data Appliance, bundled with Oracle Linux
  - Enhanced linear algebra performance: Intel’s Math Kernel Library, AMD’s Core Math Library, SUN Solaris and IBM AIX
  - Enterprise support for customers of Oracle R Enterprise, Big Data Appliance, and Oracle Linux
  - Contribute bug fixes and enhancements to open source R

- **Oracle R Enterprise**
  - Transparent access to database-resident data from R
  - Embedded R script execution through database managed R engines
  - Statistics engine

- **Oracle R Connector for Hadoop (Part of Oracle Big Data Connectors)**
  - R interface to Oracle Hadoop Cluster on BDA and non-Oracle Hadoop clusters
  - Access and manipulate data in HDFS, database, and file system
  - Write MapReduce functions using R and execute through natural R interface

- **ROracle**
  - Open source Oracle *database interface driver* for R based on OCI
  - Maintainer is Oracle – rebuilt from the ground up; many bug fixes and optimizations
How Oracle R Enterprise Works

ORE Computation Engines

- Oracle R Enterprise tightly integrates R with the database and fully manages the data operated upon by R code.
  - The database is always involved in serving up data to the R code.
  - Oracle R Enterprise runs in the Oracle Database.

- Oracle R Enterprise eliminates data movement and duplication, maintains security and minimizes latency time from raw data to new information.

- Three ORE Computation Engines
  - Oracle R Enterprise provides three different interfaces between the open-source R engine and the Oracle database:
    1. Oracle R Enterprise (ORE) Transparency Layer
    2. Oracle Statistics Engine
    3. Embedded R
Oracle Advanced Analytics

R Graphics Direct Access to Database Data

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

MPG increases over time…
Oracle Advanced Analytics

R Enterprise Compute Engines

1. **R Engine**
   - Oracle R Enterprise packages
   - Other R packages

2. **Oracle Database**
   - SQL
   - User tables
   - Open Source
   - Results
   - R

3. **R Engine**
   - Oracle R Enterprise packages
   - Other R packages

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

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

**R Engine(s) spawned by Oracle DB**
- Database can spawn multiple R engines for database-managed parallelism
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables “lights-out” execution of R scripts
Oracle Advanced Analytics Example
Use of All 3 ORE Engines Within 1 R Script

The following example illustrates use of all 3 engines from within 1 R script.

```r
m3 <- merge(mary, mar, by="ID", all.x=T) * Join */
summary(m3) * Summary */
tables(m3)

m1s3 <- m3[SUB, by=listage=AGE, gen=GEN, FUN=median])
BF <- bsplines(listage=AGE), plot=FALSE)
ore(m1, "AGE, CLASS", groups="COUNTRY, EDUCATION")

plot_set(m1[AGE=17 & in(m1, COUNTRY) & & COUNTRY IN ("USA", "France"), c(1,4,5])

x <- ore.pull(filtered=1)
library(tseries)

y <- ore(x)

m3$newcolumn = y;
library(spread)
tx <- ore행Aly(m1[,c("ARRDELAY", "DISTANCE", "newcolumn")],

function(df) {
    df[y,]
    height(ARRDELAY = DISTANCE + newcolumn, dat)
}

class(mod)

ore(parse=TRUE, na.m = TRUE)

get.BF, * object returned from the script */
```

Data preparation/analysis/joins/visualization/summarization/correlation/cross

tabulation -> completely shipped to database for execution

Local pull of filtered subset for processing by an open source package - arima in this case. Result of

processing added to database object m3 as a derived column -> Local R engine on user’s desktop

Embedded R engine invoked to build a model on the

prepared/augmented data using yet another open source package

biglm in this case
New Features
- Oracle Data Miner 4.0
- Oracle Advanced Analytics 12c
- Oracle R Enterprise 1.3
SQL Developer/Oracle Data Miner 4.0

**Preview**

- **Oracle Data Miner GUI**
  - New **Graph node**
    - Scatter, line, bar, box plots, histograms
    - `Group_by` supported
SQL Developer/Oracle Data Miner 4.0

Preview

- New Nodes in Oracle Data Miner GUI
  - New **SQL Query** node
    - Allows any form of query/transformation/statistics
    - Insert anywhere within flow
    - Allows integration of R Scripts
  - New **Predictive Query** nodes (requires 12c, more later…)
SQL Query Node to Integrate R Scripts
SQL Query Node to Integrate R Scripts
SQL Developer/Oracle Data Miner 4.0

Preview

- Deploy entire methodology as a SQL script
  - Faster deployment
New algorithms

- Expectation Maximization (EM) Clustering
  - 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)
  - Expectation Maximization algorithm automatically determines the optimal number of clusters needed to model the data.

- Principal Components Analysis (PCA)
  - Based on SVD, powerful feature extraction method use orthogonal linear projections to capture the underlying variance of the data

- Singular Value Decomposition (SVD)
  - Scales well to very large data sizes (both rows and attributes) for very large numerical data sets (for example, data from sensors such as Radio Frequency Identification)
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

- **Algorithm Enhancements**
  - Generalized Linear Models (GLM) adds automated feature selection/creation
    - With feature creation, Generalized Linear Models use non-linear terms (up to cubic terms), leading to more powerful models and increased transparency.
  - Highly scalable and performant (1,000s to 100k’s of input attributes)
  - Decision Tree algorithm now supports nested data and can be used for text mining
  - Prediction Details for all classification and regression models
  - Character data can be specified as either categorical values or text
  - Cluster details functions
    - CLUSTER_DISTANCE returns the raw distance between each row and the cluster centroid
**12c New Features**

**New Server Functionality**

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

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

```sql
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;
```
Oracle R Enterprise 1.3
New Functionality

- **Enhancements & New Functionality**
  - New ORE package for high performance in-database predictive algorithms (ORE)
    - Tighter integration with Oracle Data Mining SQL API algorithms/auto data prep/scoring, etc.
  - Exadata storage tier scoring for R models (ORE)
    - glm, glm.nb, hclust, kmeans, lm, multinom, nnet, randomForest, rpart
  - Comprehensive sampling techniques (ORE)
    - Simple, Systematic, Stratified, Cluster, Quota, Accidental
Oracle R Enterprise 1.3
New Functionality

- Enhancements & New Functionality
  - Database persistence of R objects and models
  - Interfaces to allow ORE to be embedded into applications (ORE) backend and web UI infrastructures
  - Persistence for R objects in the database
  - In-database scoring of R algorithms
  - Improved Integration with OBIEE
Oracle R Enterprise 1.3
New Functionality

- **Time Series Supported**
  - Aggregation and moving window analysis of large time series data
    - Support for DATE, TIMESTAMP, TIMESTAMP WITH TIME ZONE, and TIMESTAMP WITH LOCAL TIME ZONE Oracle data types
  - Date arithmetic, Aggregations & Percentiles, Moving window calculations
  - Out-of-the-box integration with popular R time series forecasting packages:
    - fts, tseries, xts, zoo packages, forecast and arima packages
You Can Think of OAA 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 “Verbs”**
- PREDICT
- DETECT
- CLUSTER
- CLASSIFY
- REGRESS
- PROFILE
- IDENTIFY FACTORS
- ASSOCIATE
Learn More

Send Charlie.berger@oracle.com email and I’ll send you my “fav links”

OAA Overviews:
- Link to presentation—Big Data Analytics using Oracle Advanced Analytics In-Database Option

OAA Demos
- Link to ODM Blog entry with YouTube-like recorded of OAA/ODM presentation and several “live” demos
- YouTube OAA Oracle Data Miner Market Basket Analysis preso & demo
- YouTube OAA presentation & demo on Predictive Analytics (Classification) in an Insurance use case

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 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 Virtual Machine (VM) images of Oracle Database + ODM/ODM 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