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.
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

"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

GIGABYTES OF DATA (IN BILLIONS)

Requires capability to rapidly:

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

“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

Source: IDC 2011
“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 \((\text{Attribute Importance})\)
- Predict customer behavior \((\text{Classification})\)
- Predict or estimate a value \((\text{Regression})\)
- Find profiles of targeted people or items \((\text{Decision Trees})\)
- Segment a population \((\text{Clustering})\)
- Find fraudulent or “rare events” \((\text{Anomaly Detection})\)
- Determine co-occurring items in a “baskets” \((\text{Associations})\)
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

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
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
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

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

Differentiating Features

- **Performance and Scalability**
  - Leverages power and scalability of Oracle Database.

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

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

Traditional Analytics

Data Import
Data Mining
Model “Scoring”
Data Preparation and Transformation
Data Mining Model Building
Data Prep & Transformation
Data Extraction

Oracle Advanced Analytics

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

Savings

Traditional Analytics

Data Import
Data Mining
Model “Scoring”
Data Preparation and Transformation
Data Mining Model Building
Data Prep & Transformation
Data Extraction

Oracle Advanced Analytics

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

Savings
Turkcell İletişim Hizmetleri A.Ş.
Combating Communications Fraud

Company/Background
• Industry: Communications
• Employees: 3,583
• Annual Revenue: Over $5 Billion
• First Turkish company listed on the NYSE.

Challenges/Opportunities
• Communications fraud is a major issue—anonymous prepaid cards can be used as cash vehicles—for example, to withdraw cash at ATMs
• Prepaid card fraud can result in millions of dollars lost every year
• Monitor numerous parameters for up to 10 billion daily call-data records

Solution
• Leveraged SQL for the preparation and transformation of one petabyte of uncompressed raw communications data
• Deployed Oracle Data Mining models on Oracle Exadata to identify actionable information in less time than traditional methods
• Achieved extreme data analysis speed with in-database analytics performed inside Oracle Exadata, that enabled analysts to detect fraud patterns almost immediately

Key Products
• Oracle Exadata Database Machine X2-2 HC Full Rack
• Oracle Advanced Analytics Option

Why Oracle
• Extremely fast sifting through huge data volumes
• With fraud, time is money

“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.

Future Plans
• Develop more targeted customer campaigns
• Understand call center interactions for better service
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

```sql
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;
/
```

<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
Example

Better Information for OBI EE Reports and Dashboards

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_MONTHLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```

Likelihood to respond:

Query Result

SQL | All Rows Fetched: 1 in 0 seconds

| PREDICTION_PROB... | 0.8382936507936... |
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 submissions
- **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.
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

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Predicted Voluntary Termination Reason</th>
<th>Current Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latest salary change</td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>Current grade</td>
<td></td>
<td>ProfD</td>
</tr>
<tr>
<td>Worker's stock options compared to peer</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>Time in current grade</td>
<td></td>
<td>23.32 Months</td>
</tr>
<tr>
<td>Increase in sickness over previous year</td>
<td></td>
<td>1 Days</td>
</tr>
<tr>
<td>Normal working hours</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Time since last sickness</td>
<td></td>
<td>5.91 Months</td>
</tr>
<tr>
<td>Potential profit on stock</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ratio of vested to unvested options</td>
<td></td>
<td>1%</td>
</tr>
</tbody>
</table>

Makes less likely | Makes more likely
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
- CUST_INSUR_LTV → OUTPUT_1_25 → Update Table

Transformations
- CUSTOMERS → % → Transform
- SALES → Aggregate
- Transform → Join
- SUPPLEMENTARY_DEMOGRAPHICS

Explore Data
- Explore Data

Modeling
- Regress Build
- Class Build
- Clust Build
- Assoc Build
- Model
- Model Details
- Feature Build
- Apply Text 1
- Apply
- Test
- Anomalies Model

Text
- CUSTOMER COMMENTS → Build Text → Apply 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
Exadata + Data Mining 11g Release 2
Data Mining Model ”Scoring” Pushed to Storage

- SQL predicates and OAA models are pushed to storage level for execution

For example, find the US customers likely to churn:

```
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>Problem</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 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 Detection</strong></td>
<td>One Class SVM</td>
<td>Lack examples of target field</td>
</tr>
<tr>
<td><strong>Attribute Importance</strong></td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td></td>
<td>Apriori</td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td><strong>Association Rules</strong></td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Link analysis</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 Extraction</strong></td>
<td>Nonnegative Matrix Factorization</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
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.

```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;
```
Oracle Advanced Analytics
R Graphics Direct Access to Database Data

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

MPG increases over time...
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 Enterprise Compute Engines

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

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

R Engine(s) spawned by Oracle DB
- Database can spawn multiple R engines for database-managed parallelism
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables “lights-out” execution of R scripts
Oracle 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 = relational_table_1;
m3 = relational_table_2;
m3 <- merge(mary, mary, by="ID", all.x=T)  # Join

summary(m3) = Summary  #

m3 = filter(m3, ID, list(age=AGE, gen=GEN, FUN=median))
BP <- locate(list(AGE), plot=FALSE)
ore.com(m3, "AGE, CLASS", group="COUNTRY", "EDUCATION")
filtered_set = m3[AGE==17 & in(m3,AGE) & COUNTRY IN ("USA", "France"), c(1,4,5)]

x = ore.pull(filtered_set)
library(arima)
y = arima(x)

m3 = newcolumn = y;
library(biog)
tx <- oreBiogAry(x[,c("ARRDELAY", "DISTANCE", "newcolumn")],
  function(data) {
    library(biog)
    biog(ARRDELAY = DISTANCE + newcolumn, dat)
  })

class(mod)
mod = setSPDE, na.ma = TRUE)
qres.BP, y = 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
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
1. Link to my latest OOW presentation – Digging for Gold in your DW with Oracle Advanced Analytics Option.
2. Take a Free Test Drive of Oracle Advanced Analytics (Oracle Data Miner GUI) on the Amazon Cloud.
3. Link to ODM Blog entry with YouTube-like recorded of OAA/ODM presentation and several “live” demos.
4. Link to Getting Started w/ ODM blog entry.
5. Link to New OAA/Oracle Data Mining 2-Day Instructor Led Oracle University course.
6. Link to OAA/Oracle Data Mining Oracle by Examples (free) Tutorials on OTN.
7. Link to OAA/Oracle R Enterprise (free) Tutorial Series on OTN.
8. Link to SQL Developer Days Virtual Event w/ downloadable Virtual Machine (VM) images of Oracle Database + ODM/ODMr and e-training for Hands on Labs.
9. Main OAA/Oracle Data Mining on OTN page.
10. Main Oracle Advanced Analytics Option on OTN page.

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

Learn More