Oracle Advanced Analytics
Oracle R Enterprise & Oracle Data Mining
Extending the Database into a Comprehensive Advanced Analytics Platform

Charlie Berger
Sr. Director Product Management, Data Mining and Advanced Analytics
Oracle Corporation
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.
Oracle Advanced Analytics Option—Agenda
Extending the Database into a Comprehensive Advanced Analytics Platform

• Big Data Analytics—Oracle Advanced Analytics Option
• Oracle Data Mining
  – SQL & PL/SQL focused in-database data mining and predictive analytics
• Oracle R Enterprise
  – Integrates Open Source R with the Oracle Database
Big Data Analytics

But the bigger you get, the more likely you are to be doing extensive data mining and the more likely you are to be implementing or moving towards in-database analytics.

—Jim Kobielus, a senior data management analyst with Cambridge, Mass.-based Forrester Research Inc. quote from “Customary Data Warehouse Concepts vs. Hadoop: Forrester Makes the Call”,
Mark Brunelli, Senior News Editor
This RSS Reprints Published: 11 Aug 2011
Oracle Advanced Analytics Option
Transforming the Database into a Comprehensive Advanced Analytics Platform

• Oracle Advanced Analytics Option enables companies to "bring the algorithms to the data" vs. extracting the data to specialized and expensive dedicated statistical and data mining servers

• Oracle Advanced Analytics Option includes:
  – Oracle Data Mining
    • SQL & PL/SQL focused in-database data mining and predictive analytics
  – Oracle R Enterprise
    • Integrates the Open-Source Statistical Environment R with the Oracle Database

• Data movement is eliminated or dramatically reduced while analytical and compute intensive operations are performed inside the database, where the data resides, to increase performance, reduce cycle times required to extract information from data and reduce total cost of ownership over traditional statistical and data mining environments
Oracle In-Database Advanced Analytics

Comprehensive Advanced Analytics Platform

Oracle R Enterprise
- Popular open source statistical programming language & environment
- Integrated with database for scalability
- Wide range of statistical and advanced analytical functions
- R embedded in enterprise apps & OBIEE
- Exploratory data analysis
- Extensive graphics
- Open source R (CRAN) packages
- Integrated with Hadoop for HPC

Oracle Data Mining
- Automated knowledge discovery inside the Database
- 12 in-database data mining algorithms
- Text mining
- Predictive analytics applications development environment
- Star schema and transactional data mining
- Exadata "scoring" of ODM models
- SQL Developer/Oracle Data Miner GUI

Statistics  Advanced Analytics  Data & Text Mining  Predictive Analytics
Oracle—Hardware and Software

Engineering to Work Together

Oracle is the world's most complete, open, and integrated business software and hardware systems company

- Data Warehousing, VLDB and ILM
  - Oracle Data Mining
    - 12- in-DB data mining algorithms
    - Mine star schemas, text, transactional data
    - In-DB model build & apply—Exadata scoring
    - 50+ in-DB statistical functions
  - Oracle R Enterprise
    - Run R in-DB; function push down to SQL
    - Wide library of supported in-DB statistical functions
    - Embedded R supports all R packages

Oracle has taught the RDBMS how to perform data mining, statistical analysis, adv. analytics, etc.
Oracle’s In-Database Advanced Analytics Option

Value Proposition

• **10-100x PERFORMANCE**
  - Integrated features of the Database
  - Perform analytics in-DB to eliminate data movement
  - Reduce information latency: days-wks → mins-hours

• **10x LOWER TOTAL COST OF OWNERSHIP**
  - Eliminate/minimize expensive annual usage fees associated with traditional stats/mining packages
  - Leverage Oracle DB, DW & BI technology platform
Oracle Data Mining

Building Predictive Analytics Applications

• Oracle Data Mining provides 12 powerful in-database data mining algorithms for big data analytics as a native feature of the database
  – Designed for or big data problems involving discovering patterns and relationships in large amounts of data and oftentimes making predictions based on those patterns, Oracle Data Mining allows data analysts and data miners to mine star schemas, transactional data and unstructured data stored inside the database, build predictive models and apply them to data inside the database—all without moving data.

• Oracle Data Miner help users mine their data and define, save and share advanced analytical methodologies
  – Users who prefer a Graphical User Interface can use the Oracle Data Miner extension to SQL Developer to develop, build, evaluate, share and automate analytical workflows to solve important data driven business problems.

• Developers can use the SQL APIs and PL/SQL to build applications to automate knowledge discovery
  – The Oracle Data Miner GUI generates SQL code that application developers can use to develop and deploy SQL and PL/SQL based automated predictive analytics applications that run natively inside the Oracle Database.
What is Data Mining?

- Automatically sifts through data to find hidden patterns, discover new insights, and make predictions
- Data Mining can provide valuable results:
  - Predict customer behavior *(Classification)*
  - Predict or estimate a value *(Regression)*
  - Segment a population *(Clustering)*
  - Identify factors more associated with a business problem *(Attribute Importance)*
  - Find profiles of targeted people or items *(Decision Trees)*
  - Determine important relationships and “market baskets” within the population *(Associations)*
  - Find fraudulent or “rare events” *(Anomaly Detection)*
In-Database Data Mining

Traditional Analytics

Data Import

Data Mining Model “Scoring”

Data Preparation and Transformation

Data Mining Model Building

Data Prep & Transformation

Data Extraction

Oracle Data Mining

Results

- Faster time for “Data” to “Insights”
- Lower TCO—Eliminates
  - Data Movement
  - Data Duplication
- Maintains Security

Model “Scoring”
Data remains in the Database
Embedded data preparation
Cutting edge machine learning algorithms inside the SQL kernel of Database
SQL—Most powerful language for data preparation and transformation
Data remains in the Database

Results

- Faster time for “Data” to “Insights”
- Lower TCO—Eliminates
  - Data Movement
  - Data Duplication
- Maintains Security

Savings
Oracle Data Mining and Unstructured Data

- Oracle Data Mining mines unstructured i.e. “text” data
- Include free text and comments in ODM models
- Cluster and Classify documents
- Oracle Text used to preprocess unstructured text
# Oracle Data Mining Algorithms

<table>
<thead>
<tr>
<th>Problem</th>
<th>Algorithm</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</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>Regression</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>Anomaly Detection</td>
<td>One Class SVM</td>
<td>Lack examples of target field</td>
</tr>
<tr>
<td>Attribute</td>
<td>Minimum Description Length (MDL)</td>
<td>Attribute reduction</td>
</tr>
<tr>
<td>Importance</td>
<td></td>
<td>Identify useful data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce data noise</td>
</tr>
<tr>
<td>Association</td>
<td>Apriori</td>
<td>Market basket analysis</td>
</tr>
<tr>
<td>Rules</td>
<td></td>
<td>Link analysis</td>
</tr>
<tr>
<td>Clustering</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>Feature Extraction</td>
<td>Nonnegative Matrix Factorization</td>
<td>Text analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Feature reduction</td>
</tr>
</tbody>
</table>
SQL Developer 3.0/Oracle Data Miner 11g Release 2 GUI

- Graphical User Interface for data analyst
- SQL Developer Extension (OTN download)
- Explore data—discover new insights
- Build and evaluate data mining models
- Apply predictive models
- Share analytical workflows
- Deploy SQL Apply code/scripts
Churn models to product and “profile” likely churners
Clustering analysis to discover customer segments based on behavior, demographics, plans, equipment, etc.
Market Basket Analysis to identify potential product bundless
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;

POLICYNUMBER  PERCENT_FRAUD  RNK
------------  -----------  ----
  6532         64.78        1
  2749         64.17        2
  3440         63.22        3
   654         63.10        4
 12650         62.36        5

Automated Monthly “Application!” Just add:
  Create
  View CLAIMS2_30
  As
  Select * from CLAIMS2
  Where mydate > SYSDATE – 30
In 11g Release 2, SQL predicates and Oracle Data Mining 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;
```
Real-time Prediction for a Customer

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

```sql
SELECT prediction_probability(CLAS_DT_5_2, 'Yes'
USING 7800 as bank_funds, 125 as checking_amount, 20 as credit_balance, 55 as age, 'Married' as marital_status,
250 as MONEY_MONLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```
Ability to Import/Export 3rd Party DM Models

- ODM 11g Release 2 adds ability to import 3rd party models (PMML), convert to native ODM models and score them in-DB
  - Supported models for ODM model export:
    - Decision Trees (PMML)
  - Supported algorithms for ODM model import:
    - Multiple regression models (PMML)
    - Logistic regression models (PMML)

- Benefits
  - SAS, SPSS, R, etc. data mining models can scored on Exadata
    - Imported dm models become native ODM models and inherit all ODM benefits including scoring at Exadata storage layer, 1st class objects, security, etc.
Oracle Data Mining identifies key contributors to customer churn.

Oracle Data Mining discovers different profiles of churning customers and their profile, both critical to developing a proactive response to reduce churn.
Oracle Communications Industry Data Model Example

Better Information for OBIEE Dashboards

ODM’s predictions & probabilities are available in the Database for reporting using Oracle BI EE and other tools
Exadata with Analytics and Business Intelligence
Better Together

- In-database data mining builds predictive models that predict customer behavior
- OBIEE’s integrated spatial mapping shows where

Customer “most likely” be be HIGH and VERY HIGH value customer in the future
Exadata with Analytics and Business Intelligence
Better Together

- Exadata power
- OBIEE ease-of-use

Drill-through for details about top factors that define HIGH and VERY HIGH value customers
Oracle Data Mining’s factory-installed predictive analytics show employees likely to leave, top reasons, expected performance and real-time "What if?" analysis
Predictive Analytics Applications (partial list)

- **Oracle Industry Data Models**—factory installed data mining for specific industries
  - Communications Industry Data Model—churn, segmentation, profiling, etc.
  - Retail Industry Data Model—market basket analysis, loyalty, etc.
- **Oracle Spend Classification**—auto review/real-time correction of submission mistakes
- **Oracle Fusion Human Capital Management (HCM)**
  - Predictive Workforce—Employee turnover and performance prediction
- **Oracle Fusion CRM**
  - Sales Prediction—Prediction of Sales opportunities, what to sell, amount, timing, etc.
- **Oracle Identify Management**
  - Adaptive Access Manager Real-time Security at user login
- **Oracle FMW**
  - Complex Event Processing integrated with integrated ODM models
- **Oracle ACS**
  - Predictive Incident Monitoring Service for Oracle Database customers
Learn More
R Statistical Programming Language

Open source language and environment

Used for statistical computing and graphics

Strength in easily producing publication-quality plots

Highly extensible with open source community R packages
Growing Popularity

R’s rapid adoption over several years has earned its reputation as a new statistical software standard – Rival to SAS and SPSS

While it is difficult to calculate exactly how many people use R, those most familiar with the software estimate that close to 250,000 people work with it regularly.

Typical R Approach

Statistical and advanced analyses are run and stored on the user’s laptop
What Are R’s Challenges?

1. R is memory constrained
   – R processing is single threaded - does not exploit available compute infrastructure
   – R lacks industrial strength for enterprise use cases

2. R has lacked mindshare in Enterprise market
   – R is still met with caution by the long established SAS and IBM/SPSS statistical community
     • However, major university (e.g. Yale) Statistics courses now taught in R
     • The FDA has recently shown indications for approval of new drugs for which the submission’s data analysis was performed using R
Oracle R Enterprise Approach

Data and statistical analysis are stored and run in-database.

- Same R user experience & same R clients
- Embed in operational systems
- Complements Oracle Data Mining

Highly Secure

Faster

Scalable

Open Source
What is Oracle R Enterprise?

- Integrates popular open source R statistical programming language and environment with Oracle Database 11g Release 2 for big data.
  - Designed for problems involving large amounts of data, Oracle R Enterprise integrates R with the database and enables users to run R commands on database-resident data, develop and refine R scripts, and leverage the parallelism and scalability of the database.
  - Data analysts can run the latest R open source packages and develop and operationalize R scripts for analytical applications in one step—without having to learn SQL.
  - Oracle R Enterprise performs function push down for in-database execution of base R and popular R packages.
  - Because it runs as an embedded component of the database, Oracle R Enterprise can run any R package either by function push down or via embedded R while the database manages the data served to the R engine.
What is Enterprise?

- Oracle R Enterprise brings R’s statistical functionality closer to the Oracle Database

1. Eliminate R’s memory constraint by enabling R to work directly/transparently on database objects
   - Allows R to run on very large data sets, tables, views

2. Architected for Enterprise production infrastructure
   - Automatically exploits database parallelism without requiring parallel R programming
   - Build and immediately deploy R scripts

3. Oracle R leverages the latest R algorithms and packages
   - Embedded R runs as an embedded component of the DBMS server
Architecture and Performance

- Transpareently function-ships R constructs to database via $R \rightarrow$ SQL translation
  - Data structures
  - Functions
    - Data manipulation functions (select, project, join)
    - Basic statistical functions (avg, sum, summary)
    - Advanced statistical functions (gamma, beta)

- Performs data-heavy computations in database
  - R for summary analysis and graphics

- Transparent implementation enables using wide range of R “packages” from open source community

Use Case: Using ONTIME airline data, of the 36 busiest airports, run a box-plot analysis of the best/worst airports for arrival delay?
Use Case: Using ONTIME airline data, of the 36 busiest airports, run a box-plot analysis of the best/worst airports for arrival delay?
Better Business Intelligence

*Enrich BI Dashboards with Statistics, Data Mining and Adv. Analytics*

- Ad hoc
- Explorative data analysis
- Interactive graphics
- Problem-solving

→ Oracle R Enterprise's and ODM's results become a data feed for OBIEE
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 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 Oracle Database

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

R Engine(s) spawned by Oracle DB
- Database can spawn multiple R engines for database-managed parallelism
- Efficient data transfer to spawned R engines
- Emulate map-reduce style algorithms and applications
- Enables “lights-out” execution of R scripts
How Oracle R Enterprise Works

ORE Computation Engines

1. Oracle R Enterprise (ORE) Transparency Layer
   - Traps all R commands and scripts prior to execution and looks for opportunities to function ship them to the database for native execution
   - ORE transparency layer converts R commands/scripts into SQL equivalents and thereby leverages the database as a compute engine.
How Oracle R Enterprise Works

ORE Computation Engines

2. In-Database Statistics Engine
   – Significantly extends the Oracle Database’s library of statistical functions and advanced analytical computations
   – Provides support for the complete R language and statistical functions found in Base R and selected R packages based on customer usage
     • Open source packages - written entirely in R language with only the functions for which we have implemented SQL counterparts - can be translated to execute in database.
   – Without anything visibly different to the R users, their R commands and scripts are oftentimes accelerated by a factor of 10-100x
   – ORE Functions for most common/base traditional statistical software procedures

ORE Functions
- ORE SUMMARY
- ORE FREQUENCY
- ORE CORR
- ORE UNIVARITE
- ORE CROSSTAB
- ORE RANK
- ORE SORT
- …
The following operators and functions are supported. See R documentation for syntax and semantics of these operators/function. Syntax and semantics for these items remain unchanged when used on a database mapped data type (also henceforth known as an Oracle R Enterprise data type):

- **Mathematical transformations**: abs, sign, sqrt, ceiling, floor, trunc, cummax, cummin, cumprod, cumsum, log, log10, log2, log1p, acos, acosh, asin, asinh, atan, atanh, exp, expm1, cos, cosh, sin, sinh, tan, tanh, gamma, lgamma, digamma, trigamma, round, signif, pmin, pmax, zapsmall
- **Basic statistics**: mean, summary, min, max, sum, prod, any, all, median, range, IQR, fivenum, mad, quantile, sd, var, CSS, CV, kurtosis, skewness, USS, STDERR, table, rowSums, colSums, rowMeans, colMeans
- **Arithmetic operators**: +, -, *, /, ^, %%, %/%
- **Comparison operators**: ==, >, <, !=, <=, >=
Transparency Framework + Statistics Engine

- Logical operators: &, |, xor
- Set operations: unique, %in%, match, union, intersect, setdiff, setequal, is.element
- Assignment operator forms: <-, =, ->
- String operations: tolower, toupper, casefold, toString, chatr, sub, gsub, substr, substring, paste, nchar
- Combine Data Frame: cbind, rbind, merge
- Combine vectors: c, append, stack
- Vector creation: rep, ifelse
- Subset: [, [[], $, head, tail, window, subset, Filter, na.omit, na.exclude, complete.cases
- Data order: order, rank, sort, rev
- Data reshaping: split, unlist
- Data processing: eval, with, within, transform
- Apply variants: tapply, lapply, sapply
- Aggregation: aggregate
- Regression: ore.lm() - a variant of lm()
- Anova and Boxcox: Working on output of ore.lm()
- Step-wise regression: step
- Special value checks: is.na, is.finite, is.infinite, is.nan
- Metadata functions: attributes, nrow, NROW, ncol, NCOL, nlevels, names, row, col, dimnames, dim, length, row.names, col.names, levels, reorder
- Graphics: hist, boxplot, plot, smoothScatter
- Garbage collection of temporary objects created by the transparency framework: gc
- Conversion functions: as.ore.{character, factor, integer, logical, numeric, vector}
- Test functions: is.ore.{character, factor, integer, logical, numeric, vector}
- Hypothesis testing: wilcox.test, ks.test, var.test, binom.test, chisq.test, t.test, bartlett.test
- Bessel Functions: Bessel(I,J,K,Y)
- Gamma Functions: gamma, lgamma, digamma, trigamma
- Error Functions: ERF, ERFC
- Covariance and Correlation matrices: cov, cor
- Probability, Density and Quantile Functions: Poisson, Beta, Normal,
- Chi-Squared, F, Gamma, Hyper-geometric, Probit, ProbNegative, ProbNorm, ProbT
- Matrix Operations: Matrix multiplication, Crossprod, tcrossprod, solve, backsolve, eigen, svd
How Oracle R Enterprise Works

ORE Computation Engines

3. Embedded R Engine

- The R Engine spawned by Oracle DB is also used for lights-out scheduled execution of R scripts; that is, scheduling or triggering R scripts packaged inside a PL/SQL or SQL query. Oracle R Enterprise R packages are also installed with the embedded R engine.

  - For R functions not able to be mapped to native in-database functions, Oracle R Enterprise makes “extproc” remote procedure calls to multiple R engines running on multiple database servers/nodes
  - This Oracle R Enterprise embedded layer uses the database as a data provider providing data level parallelism to R code
  - The interfaces, called embedded-layer RQ functions, pass streams of data to one or more instances of R for (parallel) row by row processing (scoring), groups of rows processing (building a model one per group) and table of rows processing (building a model

- These functions are used for “operationalizing” R code to run in production
Oracle R Enterprise Example

Illustrates Use of all 3 Engines from within 1 R Script

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

```r
# Step 1: Data Preparation
x = relational_table_1;
N = relational_table_2;
merge(x, N, by = "ID", all.x = T) -> N;
summary(N) -> N Summary;
table(N) -> N Table;
aggregate(N, by=list(age=AGE, gen=GEN), FUN=median)
BP <- boxplot(list(age=AGE), plot=FALSE)
ora.com.mi, 'AGE CLASS', group.by=c('COUNTRY', 'EDUCATION')
filtered_set = mi[AGE>17 & is.na(AGE) & COUNTRYIN ("USA", "France")]

# Step 2: Data Processing
x = ora_pull(filtered.set)
library(xts)
y = arma(ora_pull)

m3$NewColumn = y;
library(biglm)
res <- sapply(mi, function(x)
  biglm(x~ARRDELAY ~ DISTANCE + NewColumn)
)

# Step 3: Data Analysis
class(mod)
mean(res$PRED, na.rm = TRUE)
g(res, BP) -> object returned from the script
```

Data preparation/analysis/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.
R Data and Summary Statistic

**CARSTATS**

**head(CARSTATS)**
- Prints top rows of data set

**summary(CARSTATS)**
- Provides summary statistics for data set
R Graphics

R> plot(CARSTATS$weight, CARSTATS$mpg)

Heavy cars
R Graphics

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

MPG increases over time...
R Graphics

\texttt{R> plot(CARSTATS)}

- Supports Exploratory Data Analysis for Oracle data
Oracle R Enterprise ARIMA Forecasting Script

```r
year200801 <- ONTIME_S[(ONTIME_S$YEAR==2008) & (ONTIME_S$MONTH==1),]
y <- ore.pull(year200801)
gc()
delays <- tapply(y$ARRDELAY, y$DAYOFMONTH, mean, na.rm=TRUE)
delays <- ts(delays, start=1, end=31, frequency=1)
# Create a Kalman filter with the first 5 delays and predict the rest
preds <- c()
ses <- c()
# 1 step predictions
for (i in 5:length(delays))
{
  fit <- arima(delays[1:i], c(1,2,1))
  # predict 1 step into the future.
pred <- predict(fit)
preds <- c(preds, pred$pred)
ses <- c(ses, pred$se)
}
plot(5:length(delays), preds, type='l', col='green',
     ylim=range(c(preds+2*ses, preds-2*ses)), xlab="Delay of month",
     ylab="Predicted average delay (in minutes)",
     main="Average delays by day for January 2008")
lines(5:length(delays), preds+2*ses, col='red')
lines(5:length(delays), preds-2*ses, col='red')
points(5:length(delays), as.vector(delays[5:length(delays)]))
legend( 23, -8, c("Delay", "Predicted delay", "2 se confidence"),
        col=c(1, 3, 8), lty=c(0, 1, 1), pch=c(1, -1, -1), merge=TRUE)
```

[R Graphics: Device 2 (ACTIVE)]

**Average delays by day for January 2008**

- Delay
- Predicted delay
- 2 se confidence
Statistical Quality Control R Applications

• The qcc package for R:
  – Plots **Shewhart quality control charts** for continuous, attribute and count data;
  – Plots Cusum and EWMA charts for continuous data;
  – Performs **process capability analyses**;
  – Creates Pareto charts and cause-and-effect diagrams

**Process capability studies** to characterize and understand the behavior of a "process"

**Pareto (80/20 rule) analysis** to understand which few factors contribute most

Big Data Appliance + R
For Compute Intensive Operations Using R

logreg <- function(input, iterations, dims, alpha){
  plane = rep(0, dims)
  g = function(z) 1/(1 + exp(-z))
  for (i in 1:iterations) {
    z = hdfs.get(hadoop.run(input,
                      export = c(plane, g),
                      map = logisticRegressionMapper,
                      reduce = logisticRegressionReducer))
    gradient = c(z$val[1], z$val[2])
    plane = plane + alpha * gradient
  }
  plane
}
x = hdfs.push(WEBSESSIONS)
logreg(x, 10, 2, 0.05)

Function push-down
— data transformation & statistics

logreg <- function(input, iterations, dims, alpha){
  plane = rep(0, dims)
  g = function(z) 1/(1 + exp(-z))
  for (i in 1:iterations) {
    z = hdfs.get(hadoop.run(input,
                      export = c(plane, g),
                      map = logisticRegressionMapper,
                      reduce = logisticRegressionReducer))
    gradient = c(z$val[1], z$val[2])
    plane = plane + alpha * gradient
  }
  plane
}
x = hdfs.push(WEBSESSIONS)
logreg(x, 10, 2, 0.05)
**Oracle In-Database Advanced Analytics**

**Comprehensive Advanced Analytics Platform**

**Oracle R Enterprise**
- Popular open source statistical programming language & environment
- Integrated with database for scalability
- Wide range of statistical and advanced analytical functions
- R embedded in enterprise apps & OBIEE
- Exploratory data analysis
- Extensive graphics
- Open source R (CRAN) packages
- Integrated with Hadoop for HPC

**Oracle Data Mining**
- Automated knowledge discovery inside the Database
- 12 in-database data mining algorithms
- Text mining
- Predictive analytics applications development environment
- Star schema and transactional data mining
- Exadata "scoring" of ODM models
- SQL Developer/Oracle Data Miner GUI

---

**Statistics**  **Advanced Analytics**  **Data & Text Mining**  **Predictive Analytics**