Session 1: Introduction to Oracle's R Technologies

With Oracle Machine Learning

Mark Hornick, Senior Director
Oracle Machine Learning Product Management

November 2020
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, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.
Agenda

1. What is R
2. Oracle Machine Learning overview
3. Oracle R Distribution
4. ROracle Package
5. Oracle Machine Learning for Spark
6. Oracle Machine Learning for R
7. Summary
What is R?


An integrated suite of software facilities for data manipulation, calculation and graphical display.

Millions of R users worldwide:
- Widely taught in Universities
- Many Corporate Analysts and Data Scientists know and use R

Thousands of open sources packages to enhance productivity such as:
- Bioinformatics with R
- Spatial Statistics with R
- Financial Market Analysis with R
- Linear and Non Linear Modeling
Why data scientists | statisticians | data analysts use R
R is a statistics language similar to Base SAS or SPSS statistics

R environment is ..
• Powerful
• Extensible
• Graphical
• Extensive statistics
• OOTB functionality with many ‘knobs’ but smart defaults
• Ease of installation and use
• Free

http://cran.r-project.org/

Copyright © 2020 Oracle and/or its affiliates.
Analytic Pain Points

It takes too long to get my data or to get the ‘right’ data
I can’t analyze or mine all of my data – it has to be sampled
Putting analytics/predictive models and results into production is ad hoc and complex
Recoding R or other models into SQL, C, or Java takes time and is error prone
Our company is concerned about data security, backup and recovery
We need to build 10s of thousands of models fast to meet business objectives

See the blog series at https://blogs.oracle.com/R/entry/addressing_analytic_pain_points
Oracle Machine Learning
Oracle Machine Learning differentiators

Work directly with data in Database and Hadoop
Eliminate need to request extracts from IT/DBA – immediate access to database and Hadoop data
Process data where they reside – minimize or eliminate data movement

Scalability and Performance
Use parallel, distributed algorithms that scale to big data on Oracle Database
Leverage Exadata-class machines to build models on billions of rows of data

Ease of deployment
Using Oracle Database, place R, Python, and SQL scripts immediately in production (no need to recode)
Use production quality infrastructure without custom plumbing or extra complexity

Process support
Maintain and ensure data security, backup, and recovery using existing processes
Store, access, manage, and track analytics objects (models, scripts, workflows, data) in Oracle Database
Oracle Machine Learning

OML4SQL
SQL API

OML4R
R API

OML4Py*
Python API

OML Services*
Model Deployment and Management, Cognitive Text

OML Notebooks
with Apache Zeppelin on Autonomous Database

Oracle Data Miner
Oracle SQL Developer extension

OML4Spark
R API on Big Data

OML AutoML UI*
Code-free AutoML interface on Autonomous Database

* Coming soon
Oracle Machine Learning interfaces to Oracle Database

**Tool**
- Oracle Database
- OML Notebooks
- OML4SQL
- OML4Py*
- OML4R*
- Apache Zeppelin
- Python client, Jupyter Notebooks
- SQL Developer, SQL*Plus
- R client, RStudio
- SQL Developer

**Oracle Machine Learning Component**
- OML Notebooks
- OML4SQL
- OML4Py*
- OML4R*

**Data Management Platform**
- Autonomous Database
- Oracle Database
- Database Cloud Service

*Coming soon*
Oracle Machine Learning Algorithms and Analytics

**Classification**
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine (SVM)
- Explicit Semantic Analysis
  - XGBoost*

**Anomaly Detection**
- One-Class SVM
  - MSET-SPRT*

**Clustering**
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

**Time Series**
- Forecasting - Exponential Smoothing
  - Includes popular models e.g. Holt-Winters with trends, seasonality, irregularity, missing data

**Regression**
- Linear Model
- Generalized Linear Model (GLM)
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- LASSO
  - XGBoost*

**Attribute Importance**
- Minimum Description Length
- Principal Component Analysis (PCA)
- Unsupervised Pair-wise KL Div
- CUR decomposition for row & AI

**Association Rules**
- A priori/market basket

**Predictive Queries**
- Predict, cluster, detect, features

**SQL Analytics**
- SQL Windows
- SQL Patterns
- SQL Aggregates

**Feature Extraction**
- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

**Row Importance**
- CUR Decomposition

**Ranking**
- XGBoost*

**Text Mining Support**
- Algorithms support text columns
  - Tokenization and theme extraction
  - Explicit Semantic Analysis (ESA)

**Statistical Functions**
- min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

**R and Python Packages**
- Third-party R and Python Packages through Embedded Execution
- Spark MLlib algorithm integration

* New in 20c
Oracle Machine Learning Notebooks
Autonomous Database as a Data Science Platform

Collaborative UI

• Based on Apache Zeppelin
• Supports data scientists, data analysts, application developers, DBAs with SQL and Python
• Easy sharing of notebooks and templates
• Permissions, versioning, and execution scheduling

Included with Autonomous Database

• Automatically provisioned, managed, backed up
• In-database algorithms and analytics functions
• Explore and prepare, build and evaluate models, score data, deploy solutions
• Soon to be augmented with R
Oracle Machine Learning for SQL
Empower SQL users with immediate access to ML included with Oracle Database and Oracle Autonomous Database

In-database, parallelized, distributed algorithms
  • No extracting data to separate ML engine
  • Fast and scalable
  • Batch and real-time scoring
  • Explanatory prediction details
ML models as first-class database objects
  • Access control via permissions
  • Audit user actions
  • Export / import models across databases
Leverage ML across Oracle stack

SQL Interfaces
SQL*Plus
SQLDeveloper...

OML Notebooks

Oracle Database with OML

Oracle Autonomous Database

Copyright © 2020 Oracle and/or its affiliates.
Oracle Data Miner User Interface
Create analytical workflows – productivity tool for data scientists – enables citizen data scientists

SQL Developer Extension for Oracle Database on premise and DBCS
Automates typical data science steps
Easy to use drag-and-drop interface
Analytical workflows quickly defined and shared
Wide range of algorithms and data transformations
Generate SQL code for immediate deployment
Oracle Machine Learning for R and Python
Empower data scientists with open source environments

Oracle Database as HPC environment
In-database parallelized and distributed machine learning algorithms
Manage scripts and objects in Oracle Database
Integrate results into applications and dashboards via SQL or REST
OML4Py automated machine learning
Sample Use cases

Detect fraud in customer transactions, insurance claims
Identify which patients are at risk of developing certain conditions
Target the right customer with the right offer
Discover hidden customer segments
Forecast customer demand for a product or service
Find most profitable selling opportunities
Anticipate and preventing customer churn
Identify customers likely to churn and why
Security and suspicious activity detection
Understand sentiments in customer conversations
Understand influencers in social networks
Predict credit risk
Oracle’s R Technologies
Supporting R, Oracle Database, and Big Data Appliance/Hadoop

Oracle R Distribution

ROracle

Oracle Machine Learning for R
*Included with Oracle Database license and Oracle Database Cloud Service*

Oracle Machine Learning for Spark
*Component of the Big Data Connectors Software Suite and Big Data Service*
Oracle R Distribution

Ability to dynamically load
- Intel Math Kernel Library
- AMD Core Math Library
- Solaris Sun Performance Library

Oracle Support

- An Oracle-Supported Redistribution of Open Source R, now R 3.6.1
- Enhanced linear algebra performance via dynamically loaded libraries
- Improve performance at client and database for embedded R execution
- Enterprise support for customers of Oracle Advanced Analytics option, Big Data Appliance, and Oracle Linux
- Free download
- Oracle contributes bug fixes and enhancements to open source R
ORD Performance with MKL

Oracle R Distribution 3.6.1 + MKL - x64 Benchmark Results

Oracle R Distribution 3.6.1 - x64 Benchmark Results
ROracle Package
ROracle

R package enabling scalable and performant connectivity to Oracle Database

- Open source, publicly available on CRAN
- Oracle is maintainer

Oracle Database Interface (DBI) for R

- Re-implemented and optimized driver based on OCI
- Execute SQL statements from R interface
- Enables transactional behavior for insert, update, and delete
ROracle Example – enabling transactional behavior

drv <- dbDriver("Oracle")
con <- dbConnect(drv, username = "scott", password = "tiger")
dbReadTable(con, "EMP")
rs <- dbSendQuery(con, "delete from emp where deptno = 10")

dbReadTable(con, "EMP")
if(dbGetInfo(rs, what = "rowsAffected") > 1){
    warning("dubious deletion -- rolling back transaction")
    dbRollback(con)
}
dbReadTable(con, "EMP")
Oracle Machine Learning for Spark (OML4Spark)
Oracle Machine Learning for Spark
R Language API Component to Oracle Big Data Connectors

Leverage Spark 2 environment for powerful data preparation and machine learning
Use data across range of Data Lake sources
Achieve scalability and performance using full Hadoop cluster
Parallel and distributed ML algorithms from native and Spark MLlib implementations
Oracle Machine Learning for Spark
R Language API Component to Oracle Big Data Connectors

Transparency layer
• Proxy objects reference data from file system, HDFS, Hive, Impala, Spark DataFrame and JDBC sources
• Overloaded R functions translate functionality to native language, e.g., HiveQL for HIVE and Impala
• Users manipulate data via standard R syntax

Parallel, distributed machine learning algorithms
• Scalability and performance leveraging full Hadoop cluster
• Spark-based custom LM, GLM, NN, K-Means plus Spark MLlib
• Use expressive R Formula specification

Compute framework with custom R mappers/reducers
• Data-parallel and task-parallel execution
• Allows for open source CRAN packages run on Cluster Nodes
OML4Spark Performance

Logistic Regression (GLM)
Data fits in memory
• Up to 7x faster than Spark MLlib
Data cannot fit memory
• Able to solve a 10B row model
Benchmark environment
• ORAAH 2.8.0
• Big Data Appliance X7-2
• 6 Nodes, 256GB of RAM per Node

Formula: cancelled ~ distance + origin + dest + as.factor(month) + as.factor(year) + as.factor(dayofmonth) + as.factor(dayofweek) + as.factor(flightnum)
Oracle Machine Learning for R (OML4R)
Traditional R and Database Interaction

Access latency
Paradigm shift: R → SQL → R
Memory limitation – data size, call-by-value
Single threaded
Ad hoc production deployment
Issues for backup, recovery, security
Oracle Machine Learning for R
Component of Oracle Database

Use Oracle Database as HPC environment
Use in-database parallel and distributed machine learning algorithms
Manage R scripts and R objects in Oracle Database
Integrate R results into applications and dashboards via SQL
Oracle Machine Learning for R
Component of Oracle Database

Transparency layer
• Leverage proxy objects so data remains in database
• Overload R functions translating functionality to SQL
• Use standard R syntax to manipulate database data

Parallel, distributed machine learning algorithms
• Scalability and performance
• Exposes in-database algorithms from OML4SQL
• Additional R-based algorithms executing at database server

Embedded R execution
• Manage and invoke R scripts from Oracle Database
• Data-parallel, task-parallel, and non-parallel execution
• Use open source CRAN packages
Book on Oracle R Enterprise (OML4R)
Available on Amazon

Oracle R Enterprise
Harnessing the Power of R in Oracle Database: Transform Your Organization’s Big Data Into Valuable Assets
OML4R Algorithms

**Classificación**
- Decision Tree
- Logistic Regression
- Naive Bayes
- Support Vector Machine
- Random Forest

**Regressão**
- Linear Model
- Generalized Linear Model
- Multi-Layer Neural Networks
- Stepwise Linear Regression
- Support Vector Machine

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

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

**Feature Extraction**
- Nonnegative Matrix Factorization
- Principal Component Analysis
- Singular Value Decomposition
- Explicit Semantic Analysis

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

**Time Series**
- Single Exponential Smoothing
- Double Exponential Smoothing

...plus open source R packages for algorithms in combination with embedded R data- and task-parallel execution

Supports automatic data preparation, partitioned model ensembles, integrated text mining
Invoke in-database aggregation function

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

Source data is an `ore.frame` `ONTIME_S`, which resides in Oracle Database.

The `aggregate()` function has been overloaded to accept ORE frames. `aggregate()` transparently switches between code that works with standard R `data.frames` and `ore.frames`.

Returns an `ore.frame`.

R user on desktop

Client R Engine

Transparency Layer

OML4R

Oracle Database

User tables

In-db stats

select DEST, count(*) from ONTIME_S group by DEST
ore.groupApply – partitioned data flow

```r
modList <- ore.groupApply(
  X=ONTIME_S,
  INDEX=ONTIME_S$DEST,
  function(dat) {
    lm(ARRDELAY ~ DISTANCE + DEPDELAY, dat)
  }
);
summary(modList$BOS) ## return model for Boston
```

Also includes
- ore.doEval
- ore.tableApply
- ore.rowApply
- ore.indexApply
Select important predictors with ore.odmAI

In-database processing eliminates moving data
begin
--sys.rqScriptDrop('Example2')
sys.rqScriptCreate('Example2',
'function(dat, datastore_name) {
  mod <- lm(ARRDELAY ~ DISTANCE + DEPDELAY, dat)
  ore.save(mod, name=datastore_name, overwrite=TRUE)
  TRUE
}
}');
end;
/

select *
from table(rqTableEval(
  cursor(select ARRDELAY, DISTANCE, DEPDELAY
    from ontime_s),
  cursor(select 1 "ore.connect",
    'myDatastore' as "datastore_name"
    from dual),
  'XML',
  'Example2' ));

begin
--sys.rqScriptDrop('Example3')
sys.rqScriptCreate('Example3',
'function(dat, datastore_name) {
  ore.load(datastore_name)
  prd <- predict(mod, newdata=dat)
  prd[as.integer(rownames(prd))] <- prd
  res <- cbind(dat, PRED = prd)
  res}
}');
end;
/

select *
from table(rqTableEval(
  cursor(select ARRDELAY, DISTANCE, DEPDELAY
    from ontime_s
    where year = 2003
    and month = 5
    and dayofmonth = 2),
  cursor(select 1 "ore.connect",
    'myDatastore' as "datastore_name" from dual),
  'select ARRDELAY, DISTANCE, DEPDELAY, 1 PRED from ontime_s',
  'Example3'))
order by 1, 2, 3;
Special Functions
- Gamma function
- Natural logarithm of the Gamma function
- Digamma function
- Trigamma function
- Error function
- Complementary error function

Tests
- Chi-square, McNemar, Bowker
- Simple and weighted kappas
- Cochran-Mantel-Haenzel correlation
- Cramer's V
- Binomial, KS, t, F, Wilcoxon

Base SAS equivalents
- Freq, Summary, Sort
- Rank, Corr, Univariate

Density, Probability, and Quantile Functions
- Beta distribution
- Binomial distribution
- Cauchy distribution
- Chi-square distribution
- Exponential distribution
- F-distribution
- Gamma distribution
- Geometric distribution
- Log Normal distribution
- Logistic distribution

- Negative Binomial distribution
- Normal distribution
- Poisson distribution
- Sign Rank distribution
- Student's t distribution
- Uniform distribution
- Weibull distribution
- Density Function
- Probability Function
- Quantile
Oracle Machine Learning for R deployment architecture options
Summary

Oracle supports interfaces for SQL, R, Python, and a no-code UI for in-database machine learning

Oracle enables R users with advanced analytics on Big Data
  • Oracle Database
  • Big Data Appliance and Cloudera/Hortonworks clusters with Oracle Machine Learning for Spark

Oracle’s R technologies extend open source tools for Enterprise use
  • Data analysis, exploration, and machine learning
  • Simplified application development
  • Production deployment

Enables high performance, scalability, and ease of production deployment
For more information...

oracle.com/machine-learning

Oracle Machine Learning

The Oracle Machine Learning product family enables scalable data science projects. Data scientists, analysts, developers, and IT can achieve data science project goals faster while taking full advantage of the Oracle platform.

Oracle Machine Learning consists of complementary components supporting scalable machine learning algorithms for in-database and big data environments, notebook technology, SQL and R APIs, and Hadoop/Spark environments.

See also AskTOM OML Office Hours
Thank You

Mark Hornick
Oracle Machine Learning Product Management