Session 1: Introduction to Oracle's R Technologies

Oracle R Technologies

Mark Hornick
Director, Advanced Analytics and Machine Learning
mark.hornick@oracle.com

October 2018
Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Agenda

1. What is R
2. Oracle’s Advanced Analytics
3. Oracle R Distribution
4. ROracle Package
5. Oracle R Advanced Analytics for Hadoop
6. Oracle R Enterprise
7. Summary
What is R?

• R is an Open Source scripting language and environment for statistical computing and graphics [http://www.R-project.org/](http://www.R-project.org/)

• Started in 1994 as an Alternative to SAS, SPSS and other proprietary Statistical Environments

• The R environment
  – R is 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 statisticians | data analysts | data scientists 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/
Oracle’s Advanced Analytics
Oracle’s Advanced Analytics
Fastest Way to Deliver Scalable Enterprise-wide Predictive Analytics

Key Features

- Scalable in-Database + Hadoop data mining algorithms and R integration
- Powerful predictive analytics and deployment platform
- Drag and drop workflow, R and SQL APIs
- Data analysts, data scientists & developers
- Enables enterprise predictive analytics applications
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 Advanced Analytics 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 powerful engineered systems to build models on billions of rows of data or millions of models in parallel

Ease of deployment

• Using Oracle Database, place R 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’s Advanced Analytics
Multiple interfaces across platforms — SQL, R, GUI, Dashboards, Apps

**Users**

- R programmers
  - R Clients
- Data / Business Analysts
  - SQL Developer/Oracle Data Miner
- Business Analysts/Mgrs
  - OBIEE / ODV
- Domain End Users
  - Applications

**Platform**

- ORAAH
  - Parallel, distributed algorithms

**Oracle Database Enterprise Edition**

- Oracle Advanced Analytics - Database Option
  - SQL and R Integration for Scalable, Distributed, Parallel in-Database ML Execution

**Oracle Cloud**

Copyright © 2018 Oracle and/or its affiliates. All rights reserved.
Oracle’s Advanced Analytics

Polyglot Support: In-Database and Hadoop Data Management + Advanced Analytics

Oracle Database

+ Advanced Analytics Option

Oracle Database

SQL Client

SQL DM functions API
Data Miner

R Client

OAA/ORE R integration
RStudio

Big Data SQL

Oracle

Big Data Appliance

Oracle

Big Data Appliance
Oracle Data Miner GUI
Supporting both Data Scientists and “Citizen Data Scientists”

- SQL Developer Extension
- Automates many of the typical data science steps
- Easy to use drag-and-drop interface
- Quickly define analytical workflows that can be shared
- Multiple algorithms and data transformations
- Invoke R scripts via SQL node
- Generates SQL code for immediate deployment
Book on Oracle Data Miner
Available on Amazon

• Predictive Analytics Using Oracle Data Miner: Develop & Use Data Mining Models in Oracle Data Miner, SQL & PL/SQL
Oracle’s R Technologies

Supporting R, Oracle Database, and Big Data Appliance/Hadoop

• Oracle R Distribution

• ROracle

• Oracle R Enterprise

  Component of the Oracle Advanced Analytics Option to Oracle Database

• Oracle R Advanced Analytics for Hadoop

  Component of the Big Data Connectors Software Suite

Software available to R Community for free
Oracle R Distribution
Oracle R Distribution

- An Oracle-Supported Redistribution of Open Source R, now R 3.3.0
- 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

Ability to dynamically load

Intel Math Kernel Library
AMD Core Math Library
Solaris Sun Performance Library

Oracle Support
ORD Performance with MKL

Oracle R Distribution 3.3.0 + MKL - x64 Benchmark Results

Oracle R Distribution 3.3.0 - 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 R Advanced Analytics for Hadoop
Oracle R Advanced Analytics for Hadoop (ORAAH) on Hadoop Cluster

R interface to HQL Basic Statistics, Data Prep, Joins and View creation

Parallel, distributed algorithms:
- ORAAH Spark algorithms: Deep Neural, GLM, LM
- Spark MLlib algorithms: LM, GLM, LASSO, Ridge Regression, Decision Trees, Random Forests, SVM, k-Means, PCA

Use of Open-source R packages via custom R Mappers / Reducers

R Client

Oracle Database with Advanced Analytics option

SQL Client

SQL Developer

Other SQL Apps

R Analytics

Oracle R Advanced Analytics for Hadoop
<table>
<thead>
<tr>
<th>Classification</th>
<th>Regression</th>
<th>Feature Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM ORAAH</td>
<td>MLP Neural Networks ORAAH</td>
<td>Non-negative Matrix Factorization</td>
</tr>
<tr>
<td>Logistic Regression ORAAH</td>
<td>LASSO</td>
<td>Collaborative Filtering (LMF)</td>
</tr>
<tr>
<td>Logistic Regression Spark MLlib</td>
<td>Ridge Regression Spark MLlib</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>Random Forests Spark MLlib</td>
<td>Support Vector Machines Spark MLlib</td>
<td></td>
</tr>
<tr>
<td>Decision Trees Spark MLlib</td>
<td>Random Forest Spark MLlib</td>
<td></td>
</tr>
<tr>
<td>Support Vector Machines Spark MLlib</td>
<td>Linear Regression Spark MLlib</td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>Basic Statistics</td>
<td>Attribute Importance</td>
</tr>
<tr>
<td>Hierarchical k-Means Spark MLlib</td>
<td>Correlation/Covariance Spark MLlib</td>
<td>Principal Components Analysis</td>
</tr>
<tr>
<td>Hierarchical k-Means Spark MLlib</td>
<td></td>
<td>Principal Components Analysis</td>
</tr>
<tr>
<td>Gaussian Mixture Models Spark MLlib</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>Clustering</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Hierarchical-ELM (Oracle’s MPI/Spark-based)</td>
<td>Hierarchical k-Means (Spark MLlib)</td>
<td></td>
</tr>
<tr>
<td>Extreme Learning Machines (Oracle’s MPI/Spark-based)</td>
<td>Gaussian Mixture Models (Spark MLlib)</td>
<td></td>
</tr>
<tr>
<td>Multi-Layer Neural Nets (Oracle’s Spark-based)</td>
<td>Hierarchical k-Means (also available in MapReduce)</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (Oracle’s Spark-based)</td>
<td>Feature Extraction &amp; Creation</td>
<td></td>
</tr>
<tr>
<td>Gradient Boosted Trees (Spark MLlib)</td>
<td>Distributed Stochastic PCA (Oracle’s MPI/Spark-based)</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (Spark MLlib)</td>
<td>Distributed Stochastic SVD (Oracle’s MPI/Spark-based)</td>
<td></td>
</tr>
<tr>
<td>Decision Trees (Spark MLlib)</td>
<td>Principal Component Analysis (Spark MLlib)</td>
<td></td>
</tr>
<tr>
<td>Random Forest (Spark MLlib)</td>
<td>Nonnegative Matrix Factorization (Map-Red)</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Low Rank Matrix Factorization (Map-Red)</td>
<td></td>
</tr>
</tbody>
</table>

- Multi-Layer Neural Nets (Oracle’s Spark-based)
- Linear Regression Model (Oracle’s Spark-based)
- Gradient Boosted Trees (Spark MLlib)
- Linear Regression Model (Spark MLlib)
- Support Vector Machine (SVM) (Spark MLlib)
- LASSO (Spark MLlib)
- Ridge Regression (Spark MLlib)
- Random Forest (Spark MLlib)
- Decision Trees (Spark MLlib)

### Transparency Functions with IMPALA and HIVE
- Aggregations, Table Joins, summarization
- Variable Creation, Push & Pull data from IMPALA and HIVE
- Ability to push and pull data from Oracle Database
- JDBC Driver interface - build Spark DataFrames for ORAAH

### Open Source R Algorithms
- Ability to run any R package via our hadoop.run function in Map-Reduce mode
Oracle R Enterprise
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

R script cron job

Flat Files

Database

RODBC / RJDBC / ROracle

read

export

Flat Files

extract / export

load

SQL

Copyright © 2018 Oracle and/or its affiliates. All rights reserved.
Oracle R Enterprise

*Oracle Advanced Analytics Option to 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
OAA / Oracle R Enterprise

• 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 Oracle Data Mining
  – Additional R-based algorithms executing at database server

• Embedded R execution
  – Manage and invoke R scripts in Oracle Database
  – Data-parallel, task-parallel, and non-parallel execution
  – Use open source CRAN packages
Book on Oracle R Enterprise

Available on Amazon

- Oracle R Enterprise Harnessing the Power of R in Oracle Database: Transform Your Organization’s Big Data Into Valuable Assets
## Predictive Analytics algorithms in-Database

### Classification
- Decision Tree
- Logistic Regression
- Naïve Bayes
- Support Vector Machine
- Random Forest

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

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

### Attribute Importance
- Minimum Description Length

### Anomaly Detection
- 1 Class Support Vector Machine

### Market Basket Analysis
- Apriori – Association Rules

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

### Time Series
- Single Exponential Smoothing
- Double Exponential Smoothing

* Database 12.2 only
Invoke in-database aggregation function

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

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

class(aggdata)

head(aggdata)
```

R user on desktop

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

[1] "ore.frame"
attr(,"package")
[1] "OREbase"

R> class(aggdata)

[1] "ore.frame"

R> head(aggdata)

  Group.1 x
0   ABE 237
1   ABI  34
2  ABO 1357
3   ABY  10
4   ACK  3
5   ACT  33
```
ore.groupApply – partitioned data flow

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 – eliminates moving data
Embedded R Execution – SQL Interface
For model build and batch scoring

```r
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'));
```

```r
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;
```
Oracle R Enterprise Deployment Architecture

- Web browser
- Third-Party RStudio Server with ORE Client Open Source R or ORD
- OBIEE Oracle Data Visualization
- ORE Client Open Source R or ORD
- BDA / Hadoop
- Big Data SQL
- Oracle Database OAA (ORE) ORD
- R Script Repository
- R Object Datastore
Oracle is a founding member of the R Consortium

- **R Consortium Central mission** – work with and provide support to the R Foundation and R Community including key organizations developing, maintaining, distributing and using R software through the identification, development and implementation of infrastructure projects

- Enable the R user community to grow without disrupting R language development or the work of the R Foundation

- Organized under an open source governance and foundation model
  - Consists of Board of Directors, Infrastructure Steering Committee, other committees as needed
  - **Linux Foundation** provides backend operational support, guidance on operational practices from similar projects, and program management resources to help the R Consortium achieve maximum impact.

- See [https://www.r-consortium.org](https://www.r-consortium.org)
Summary

• Oracle supports interfaces for SQL, R, and GUI users for in-database Advanced Analytics

• Oracle enables R users with advanced analytics on Big Data
  – Oracle Database with Oracle Advanced Analytics – Oracle R Enterprise
  – Big Data Appliance and Cloudera/Hortonworks clusters with Oracle R Advanced Analytics for Hadoop

• 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
To Learn More about Oracle’s R Technologies...

http://oracle.com/goto/R

R Technologies from Oracle
Bringing the Power of R to the Enterprise