Using R for Big Data Advanced Analytics and Machine Learning Hands-On Lab

Using Oracle R Enterprise

Mark Hornick
Marcos Arancibia
Oracle Advanced Analytics

February 1, 2017

**BIWA SUMMIT 2017
WITH SPATIAL SUMMIT**

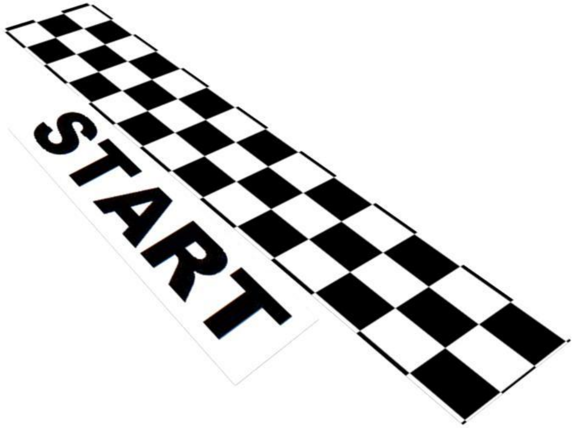
THE Big Data + Analytics + Spatial + Cloud + IoT + Everything Cool User Conference
January 31 - February 2, 2017

Copyright © 2017 Oracle and/or its affiliates. All rights reserved. |

Safe Harbor Statement

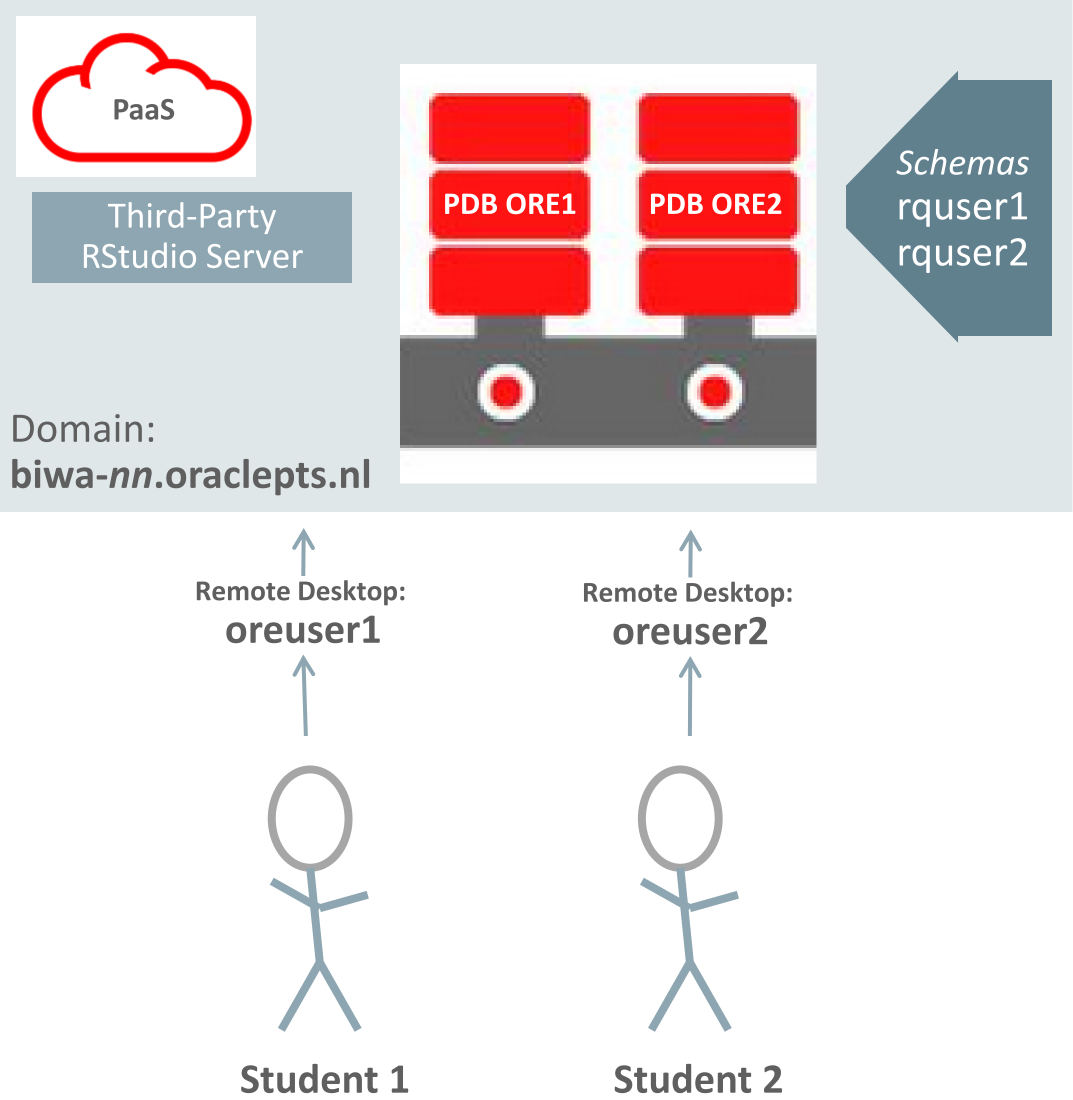
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development,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

ORACLE®

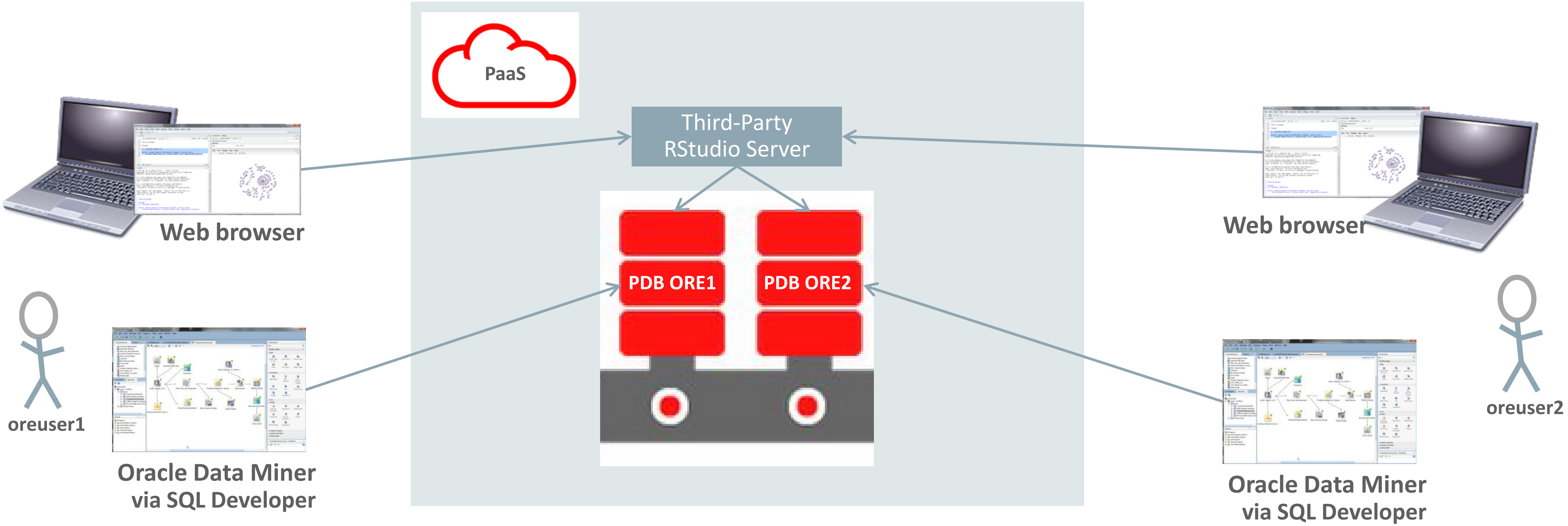


Connect to the ORE HOL Instance

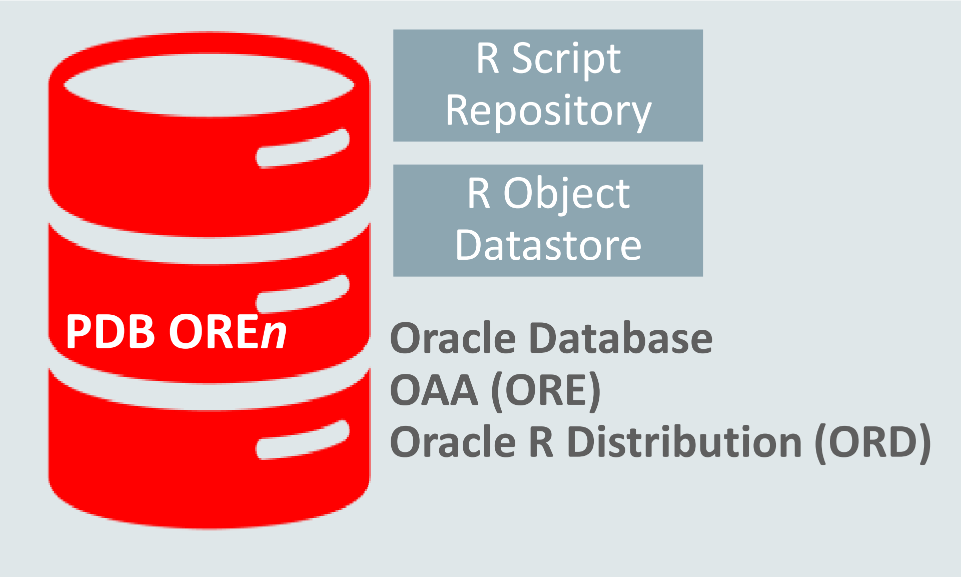
Oracle R Enterprise Cloud Deployment Architecture



Oracle R Enterprise Cloud Deployment Architecture


PaaS
Third-Party RStudio Server
PDB ORE1
PDB ORE2
Web browser
oreuser1
Oracle Data Miner via SQL Developer
Web browser
oreuser2
Oracle Data Miner via SQL Developer

Oracle R Enterprise Cloud Deployment Architecture



Oracle Database
OAA (ORE)
Oracle R Distribution (ORD)

MacOS users

- Install from the Apple App Store, not Microsoft website (old version)
- https://itunes.apple.com/us/app/microsoft-remote-desktop

Domains

Use the domain from the signup sheet

- biwa-***nn***.oraclepts.nl
- Login as oreuser1

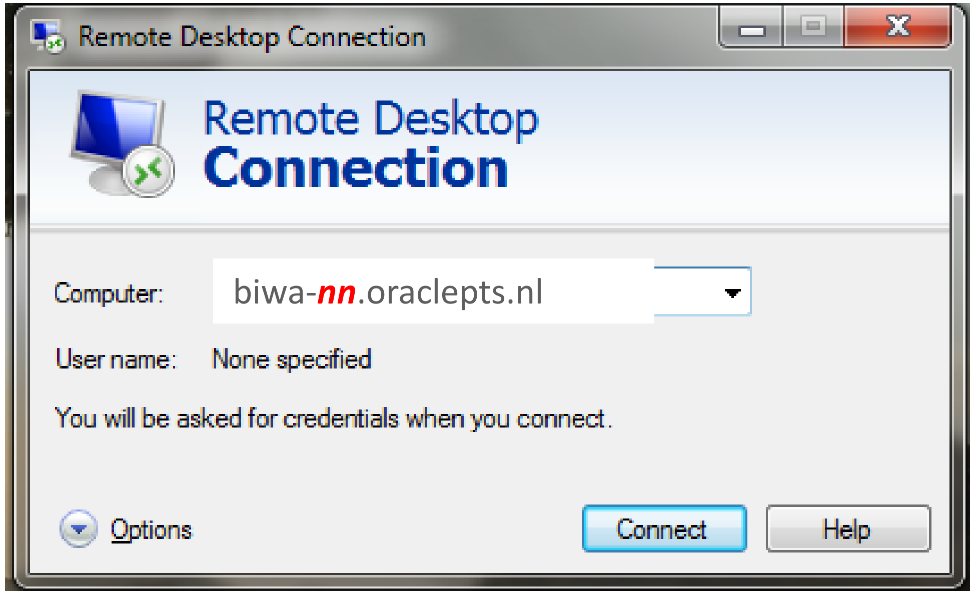
biwa-07.oraclepts.nl
biwa-08.oraclepts.nl
biwa-09.oraclepts.nl
biwa-10.oraclepts.nl
biwa-11.oraclepts.nl
biwa-12.oraclepts.nl
biwa-13.oraclepts.nl
biwa-14.oraclepts.nl
biwa-16.oraclepts.nl
biwa-17.oraclepts.nl
biwa-18.oraclepts.nl
biwa-19.oraclepts.nl
biwa-20.oraclepts.nl
biwa-21.oraclepts.nl
biwa-22.oraclepts.nl
biwa-23.oraclepts.nl
biwa-24.oraclepts.nl
biwa-25.oraclepts.nl
biwa-26.oraclepts.nl
biwa-27.oraclepts.nl
biwa-28.oraclepts.nl
biwa-29.oraclepts.nl
biwa-30.oraclepts.nl
biwa-31.oraclepts.nl
biwa-32.oraclepts.nl
biwa-33.oraclepts.nl
biwa-34.oraclepts.nl
biwa-35.oraclepts.nl
biwa-36.oraclepts.nl
biwa-37.oraclepts.nl
biwa-38.oraclepts.nl
biwa-39.oraclepts.nl
biwa-40.oraclepts.nl

biwa-48.oraclepts.nl
biwa-49.oraclepts.nl
biwa-50.oraclepts.nl
biwa-51.oraclepts.nl
biwa-52.oraclepts.nl
biwa-53.oraclepts.nl
biwa-54.oraclepts.nl
biwa-57.oraclepts.nl
biwa-58.oraclepts.nl
biwa-60.oraclepts.nl
biwa-61.oraclepts.nl
biwa-65.oraclepts.nl
biwa-66.oraclepts.nl
biwa-71.oraclepts.nl
biwa-74.oraclepts.nl
biwa-76.oraclepts.nl
biwa-77.oraclepts.nl
biwa-78.oraclepts.nl
biwa-79.oraclepts.nl
biwa-80.oraclepts.nl
biwa-81.oraclepts.nl
biwa-82.oraclepts.nl
biwa-83.oraclepts.nl
biwa-84.oraclepts.nl
biwa-85.oraclepts.nl
biwa-86.oraclepts.nl
biwa-87.oraclepts.nl
biwa-88.oraclepts.nl
biwa-89.oraclepts.nl
biwa-90.oraclepts.nl
biwa-91.oraclepts.nl
biwa-92.oraclepts.nl
biwa-93.oraclepts.nl

Connect to Remote Desktop

Student Environment

- Use the computer domain provided
- Login and Password
  - **oreuser1**   [or **oreuser2]**
  - Biwa2017



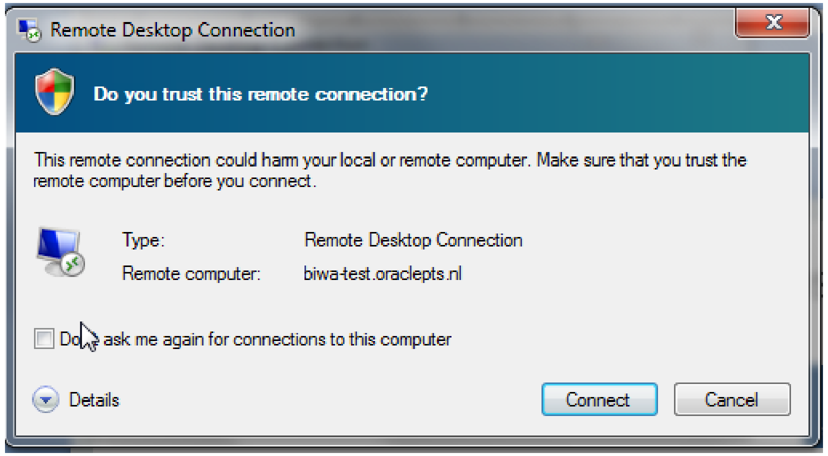
Remote Desktop Connection

Remote Desktop **Connection**

Computer: biwa-***nn***.oraclepts.nl

User name: None specified

You will be asked for credentials when you connect.



Remote Desktop Connection

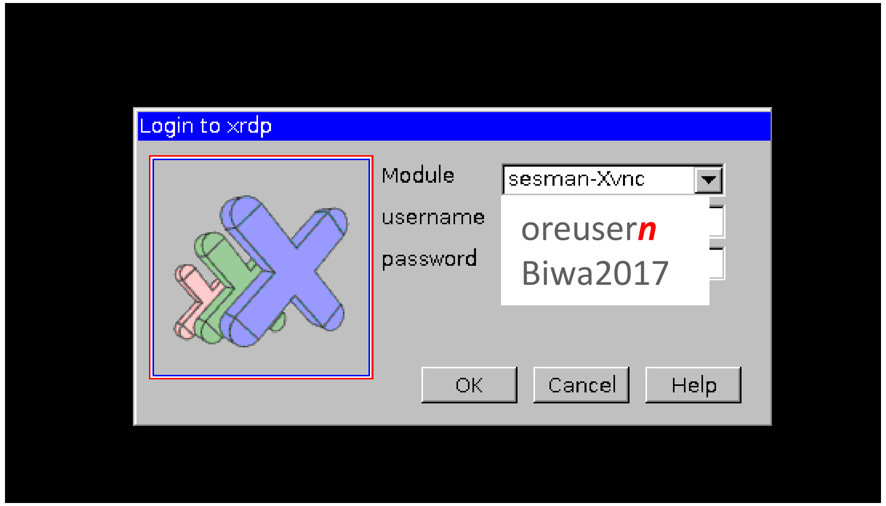
Do you trust this remote connection?

This remote connection could harm your local or remote computer. Make sure that you trust the remote computer before you connect.

Type: Remote Desktop Connection

Remote computer: biwa-test.oraclepts.nl

□ Don't ask me again for connections to this computer

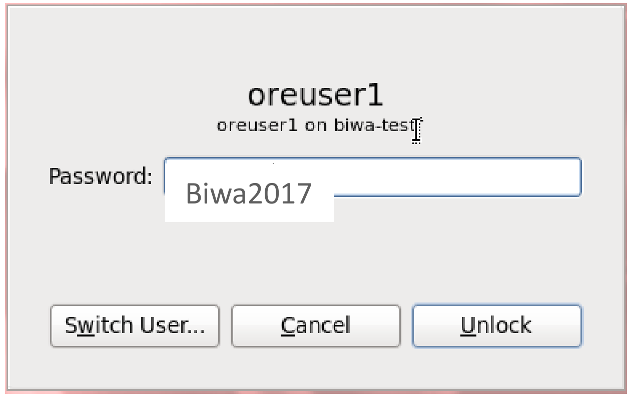


Login to xrdp

Module: sesman-Xvnc

username: oreuser***n***

password: Biwa2017



oreuser1

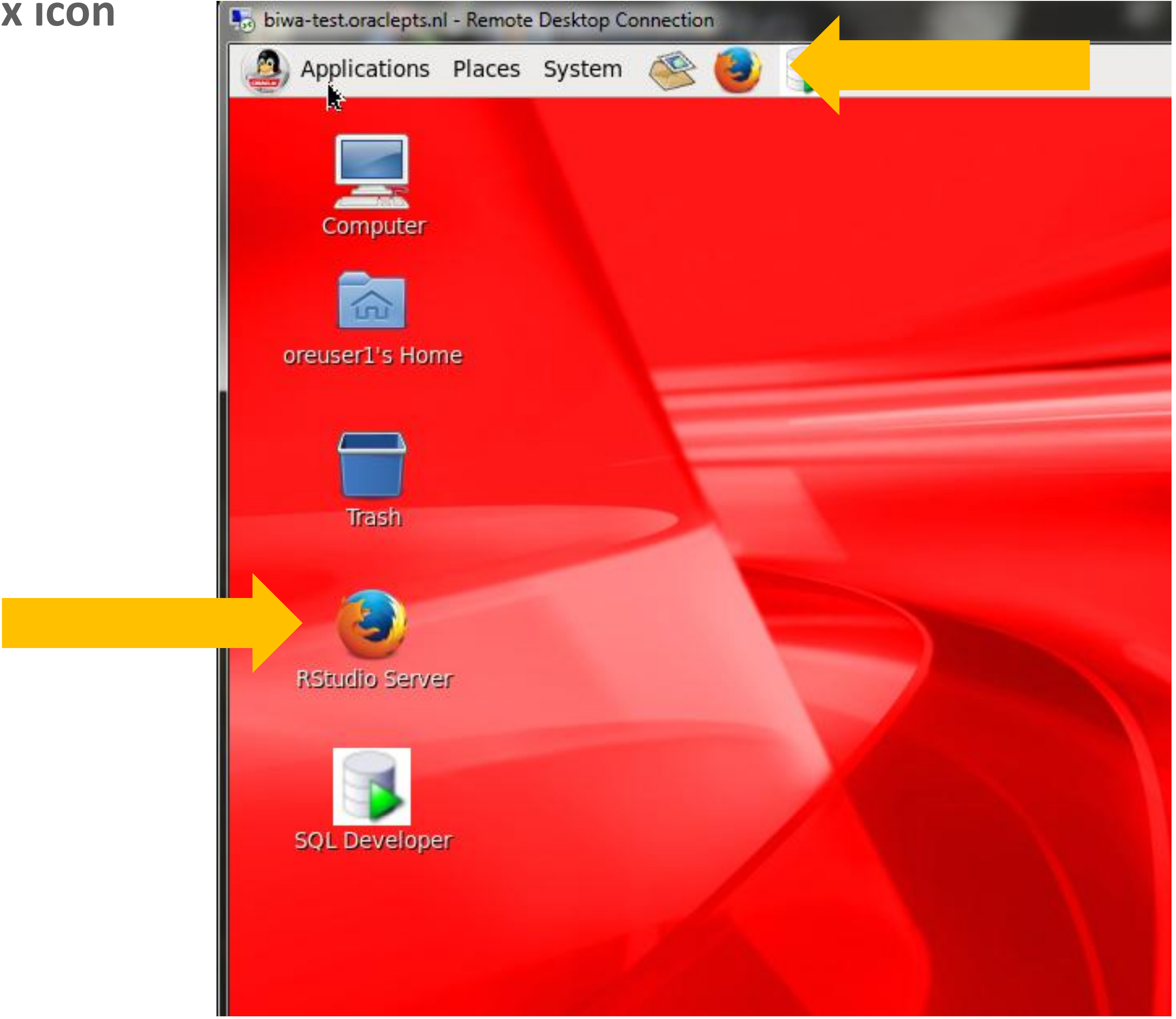
oreuser1 on biwa-test

Password: Biwa2017

Double click on “RStudio Server” Firefox icon

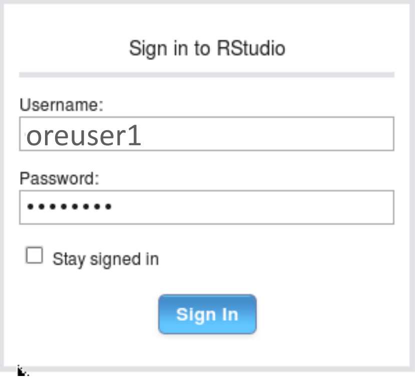
[or]

Single click the firefox icon at the top.



biwa-test.oraclepts.nl - Remote Desktop Connection

Steps to connect to ORE HOL Instance and set up
RStudio environment to access Oracle R Enterprise



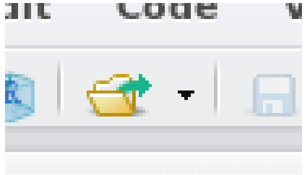
Sign in to RStudio

Username: oreuser1

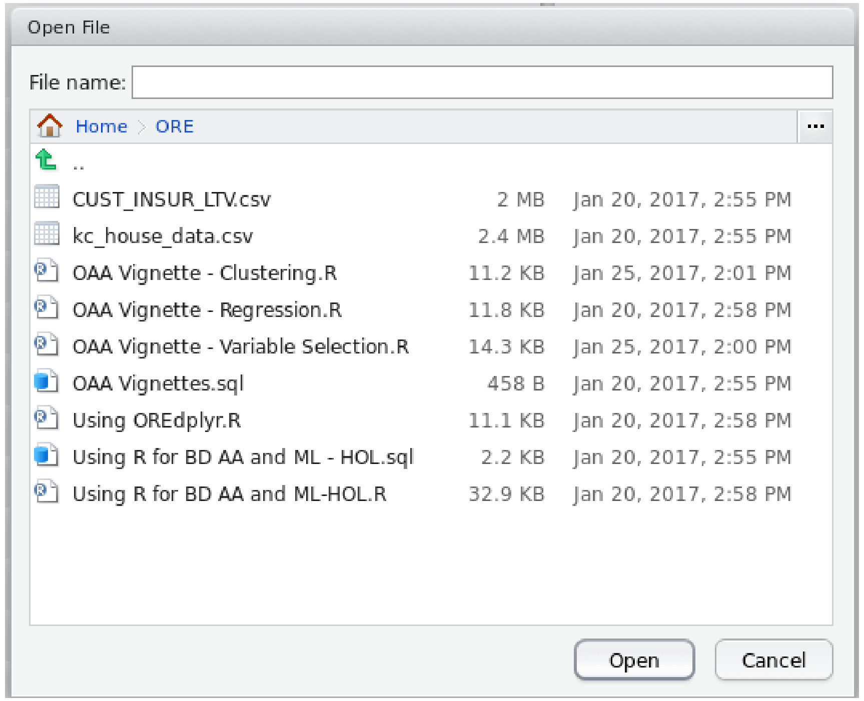
Password: ••••••••

- Sign in with user ‘oreuser1’ [or] ‘oreuser2’ and password ‘Biwa2017’

- Click the open file icon



- Click the ‘ORE’ folder
- Select the file “Oracle Vignette – Variable Selection”



Open File

File name:

Home › ORE

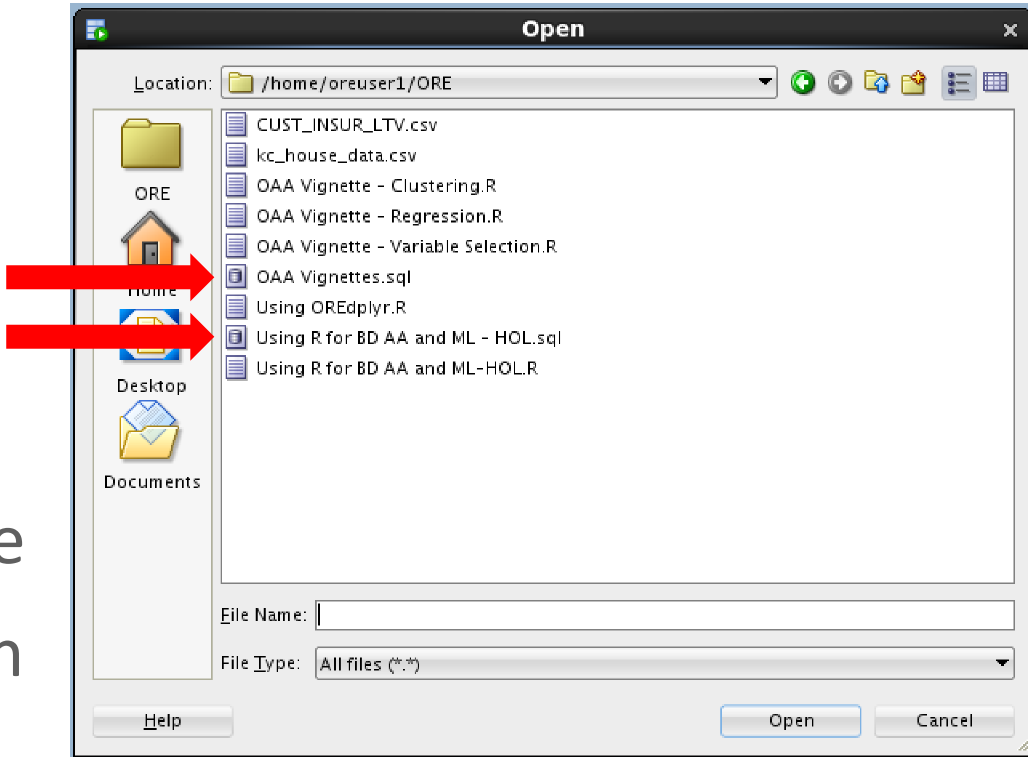
..

CUST_INSUR_LTV.csv	2 MB	Jan 20, 2017, 2:55 PM
kc_house_data.csv	2.4 MB	Jan 20, 2017, 2:55 PM
OAA Vignette - Clustering.R	11.2 KB	Jan 25, 2017, 2:01 PM
OAA Vignette - Regression.R	11.8 KB	Jan 20, 2017, 2:58 PM
OAA Vignette - Variable Selection.R	14.3 KB	Jan 25, 2017, 2:00 PM
OAA Vignettes.sql	458 B	Jan 20, 2017, 2:55 PM
Using OREdplyr.R	11.1 KB	Jan 20, 2017, 2:58 PM
Using R for BD AA and ML - HOL.sql	2.2 KB	Jan 20, 2017, 2:55 PM
Using R for BD AA and ML-HOL.R	32.9 KB	Jan 20, 2017, 2:58 PM

Steps to connect to ORE HOL Instance and set up

SQL Developer for Oracle Data Miner

- Click SQL Developer icon at top
- Click File->Open, or this icon , click 'ORE folder, and select the .sql file
- Go back to Firefox and the RStudio interface
- In RStudio, click Run or CTRL-Enter to run one line or selected set of lines



Location: /home/oreuser1/ORE

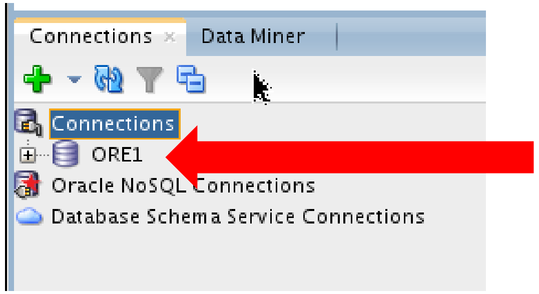
- CUST_INSUR_LTV.csv
- kc_house_data.csv
- OAA Vignette - Clustering.R
- OAA Vignette - Regression.R
- OAA Vignette - Variable Selection.R
- OAA Vignettes.sql
- Using OREdplyr.R
- Using R for BD AA and ML - HOL.sql
- Using R for BD AA and ML-HOL.R

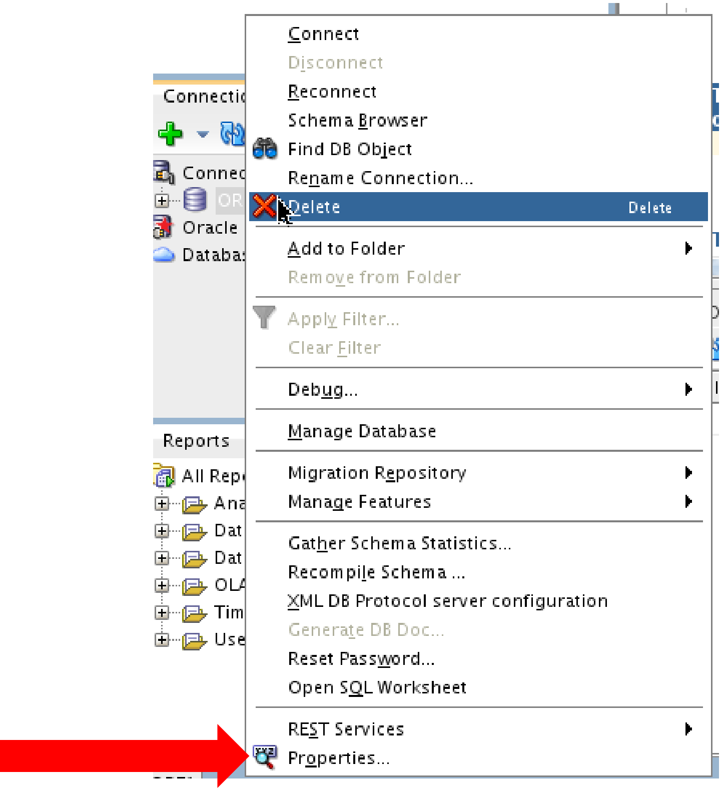
File Name:

File Type: All files (*.*)

In SQL Developer

Go to ‘Connections’
right lick “ORE*n*”
Select ‘Properties’

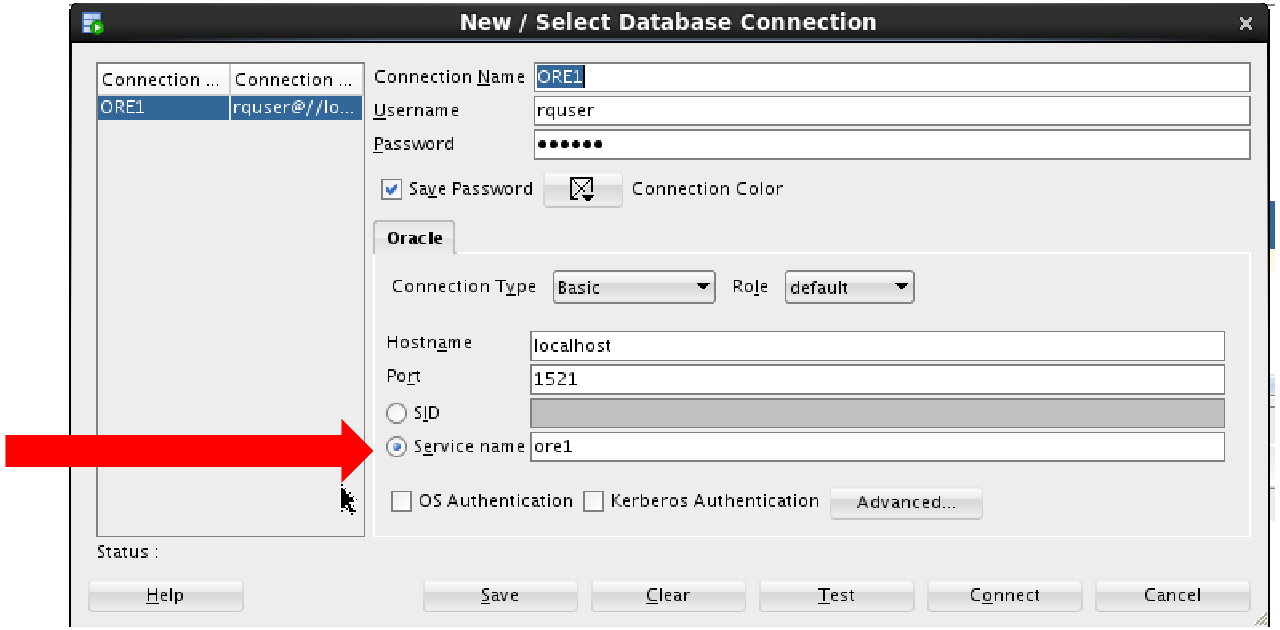




In SQL Developer

REMOVE all text after
ore*n*

i.e., **.lab14.oraclecloud.internal**



New / Select Database Connection

Connection ...	Connection ...
ORE1	rquser@//lo...

Connection Name: ORE1

Username: rquser

Password: ••••••

☑ Save Password

Connection Color

Oracle

Connection Type: Basic

Role: default

Hostname: localhost

Port: 1521

○ SID

● Service name: ore1

□ OS Authentication □ Kerberos Authentication

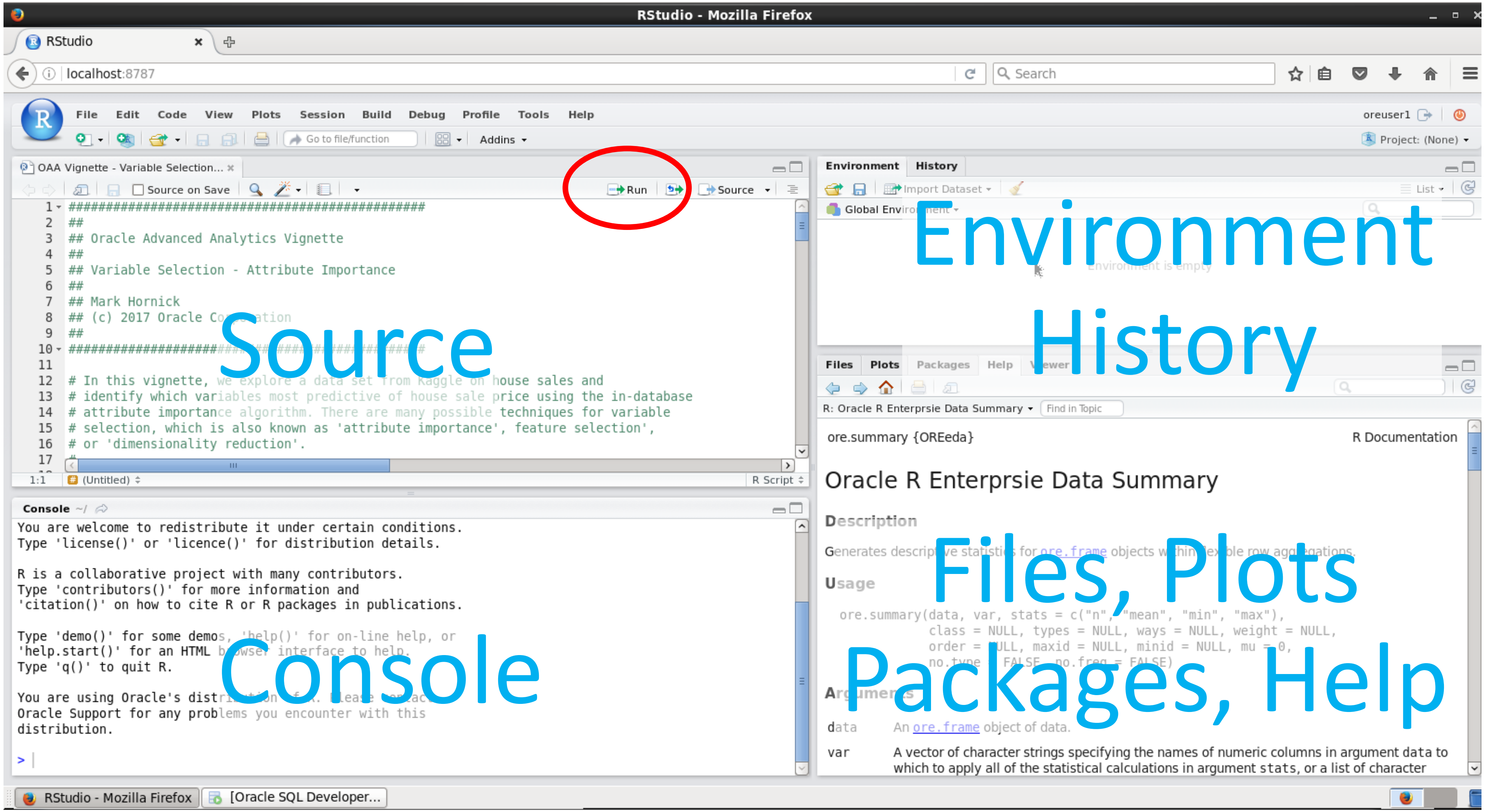
Status :

ORACLE

Back to RStudio

- Go back to Firefox and the RStudio interface
- In RStudio, click or CTRL-Enter to run one line or selected set of lines

Back to Firefox and R



```
################################################
##
## Oracle Advanced Analytics Vignette
##
## Variable Selection - Attribute Importance
##
## Mark Hornick
## (c) 2017 Oracle Corporation
##
################################################

# In this vignette, we explore a data set from Kaggle on house sales and
# identify which variables most predictive of house sale price using the in-database
# attribute importance algorithm. There are many possible techniques for variable
# selection, which is also known as 'attribute importance', feature selection',
# or 'dimensionality reduction'.
#
```

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

You are using Oracle's distribution of R. Please contact
Oracle Support for any problems you encounter with this
distribution.

>
```

ore.summary {OREeda} R Documentation

Oracle R Enterprsie Data Summary

Description

Generates descriptive statistics for ore.frame objects within flexible row aggregations.

Usage

```
ore.summary(data, var, stats = c("n", "mean", "min", "max"),
            class = NULL, types = NULL, ways = NULL, weight = NULL,
            order = NULL, maxid = NULL, minid = NULL, mu = 0,
            no.type = FALSE, no.freq = FALSE)
```

Arguments

data	An ore.frame object of data.
var	A vector of character strings specifying the names of numeric columns in argument data to which to apply all of the statistical calculations in argument stats, or a list of character

Source

Environment History

Console

Files, Plots Packages, Help

Hands-on Lab Format and Content

- Follow along with instructor through first script, or go at your own pace
  - OAA Vignette – Variable Selection using Attribute Importance
  - OAA Vignette – Clustering
  - OAA Vignette – Regression
  - OREdplyr
  - Using R for Big Data AA and ML

Note: we will not cover all of these as a group

- Click “run” on each line, or group of lines to execute
- Explore beyond the script if you’re comfortable with R
- Get online doc for any function you need help with ***?functionName***

OAA Vignette – Variable Selection using Attribute Importance

- Learn the basics of interacting with R and ORE
- Create HOUSE dataset and table from a file
- Explore and prepare data using in-database execution from ORE Transparency Layer
- Visualize data using both overloaded functions and CRAN package, e.g., ggplot2
- Perform variable selection using in-database Attribute Importance function
- Use embedded R execution from R and SQL with ORE datastore and R script repository
- *Attribute importance selection can be a step before classification or regression model building*

OAA Vignette – Clustering

- Generate simple 2D data with 3 clusters in R and push to Oracle Database
- Build k-Means and O-Cluster models, assign clusters, and visualize results
- Use auto data set from ISLR package
- Build k-Means clustering model , assign clusters, and visualize results
- Explore clusters with a few statistics and 2D ggplot2 visualization
- Visualize clusters in 3D using plot3D
- Use Oracle Data Miner to build multiple models and visualize results
- Generate plots using ORE embedded R execution from both R and SQL

OAA Vignette – Regression

- Create HOUSE dataset and table from a file
- Explore and prepare data using in-database execution from ORE Transparency Layer
- Sample data into train and test sets using ORE row indexing
- Build a variety of models and score data using:
  - R: lm
  - ORE: ore.lm, ore.odmSVM, ore.odmGLM, ore.neural
- Use ORE embedded R execution to build one model per zipcode and store in datastore

Using OREdplyr

- Use the package nycflights13 and mtcars datasets and create database tables
- Explore basic operations of the overloaded dplyr functions in OREdplyr
  - These use the same API as dplyr, but accept ore.frame objects for in-database execution
  - Functions:
    select, rename, filter, arrange, distinct, mutate, transmutate, summarise, slice, sample_n, tally
  - Stacking operations
  - Groupling with group_by
  - Chaining
- Contrast non-standard evaluation and standard evaluation
- Table joins
  - Functions: inner_join, left_join, right_join, full_join

Using R for Big Data AA and Machine Learning *(advanced - long)*

- Broader range of functionality of ORE
- Loading data and accessing across database schemas – granting access
- Accessing shared datastores
- Exploring data – statistics and visualization
- Preparing data – recode, bin, normalize, outlier treatment
- Sampling

- Model building and scoring
- Embedded R Execution – parallel building on partitioned data
- Viewing models in Oracle Data Miner
- In-database scoring using R models
- Solution deployment using embedded R execution with the R and SQL interfaces
- Sharing R scripts

ORE Introduction

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

Scaling R to Big Data

Immediate access to database and Hadoop data from R

- Eliminate need to request extracts from IT/DBA
- Process data where they reside – minimize or eliminate data movement – through data.frame proxies

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 from R

Ease of deployment

- Using Oracle Database, place R scripts immediately in production (no need to recode) via SQL
- Use production quality infrastructure without custom plumbing or extra complexity

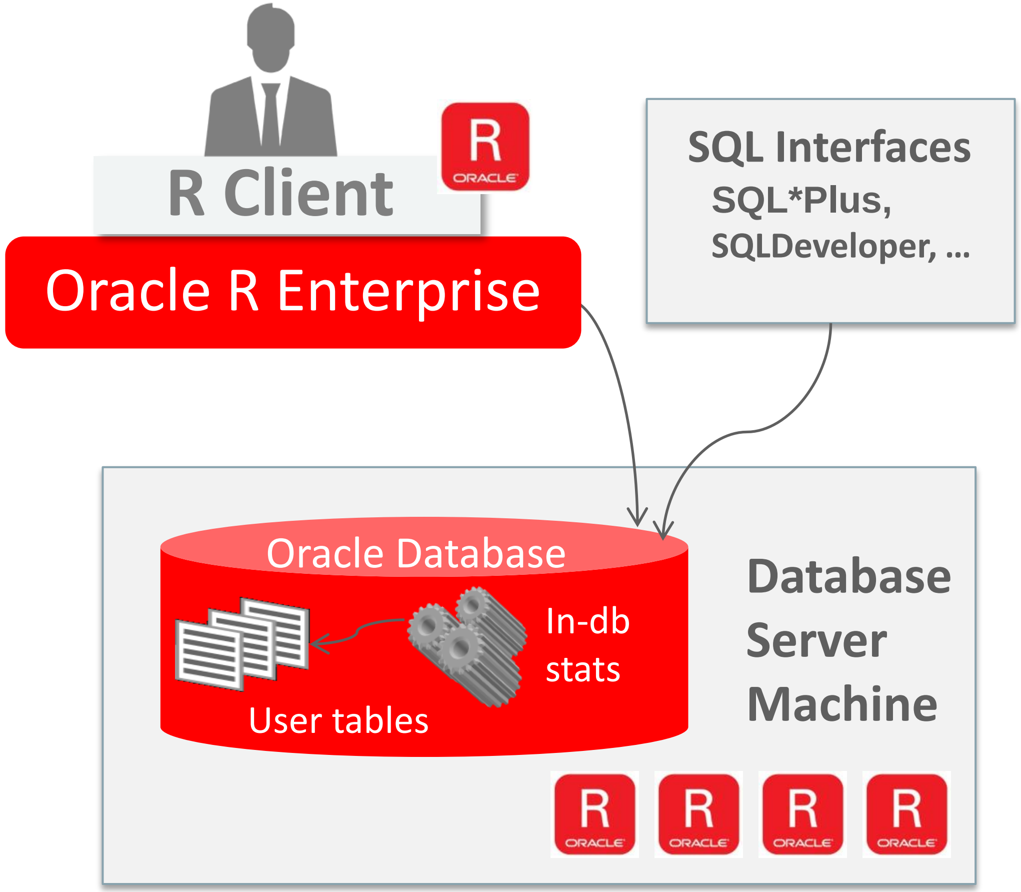
Process support

- 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 R Enterprise

Part of Oracle Advanced Analytics option to Oracle Database

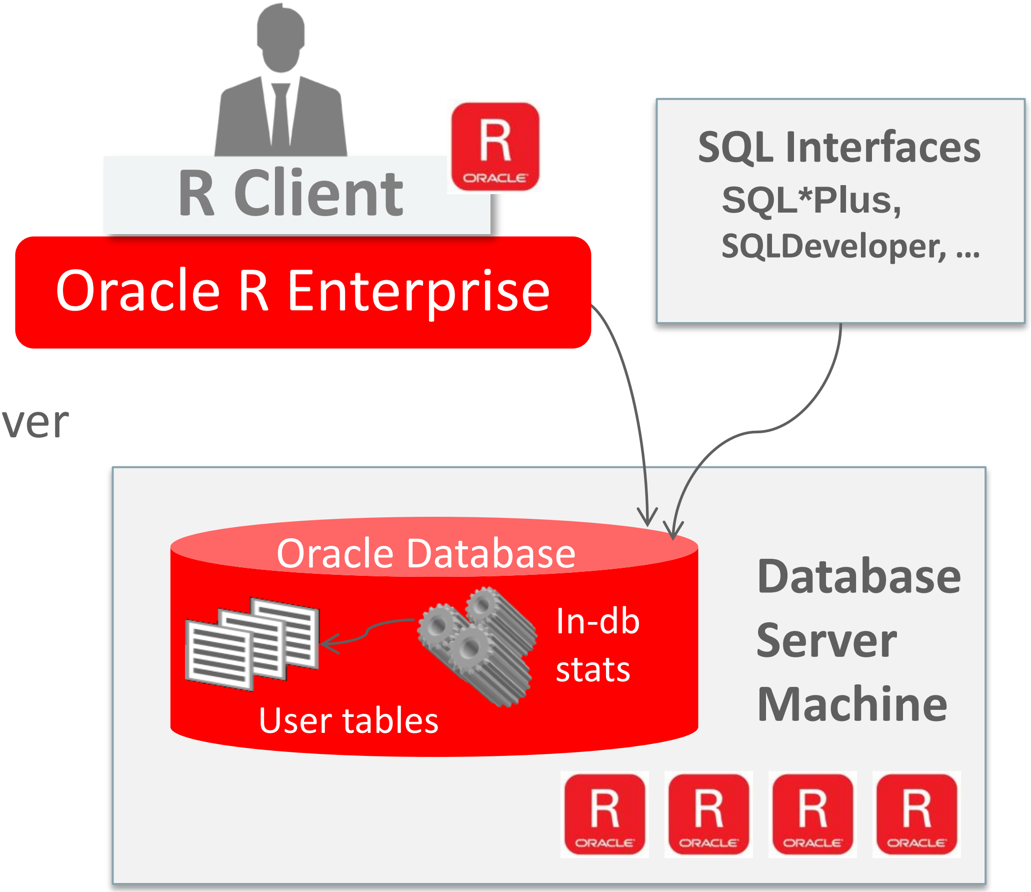
- Use Oracle Database as HPC environment
- Use in-database parallel and distributed machine learning algorithms
- Manage R scripts and R objects in Oracle Database
- Integrate R results into applications and dashboards via SQL



Oracle R Enterprise

Part of Oracle Advanced Analytics option to Oracle Database

- Transparency layer
  - Leverage proxy objects (ore.frames) - data remains in the database
  - Overload R functions that translate functionality to SQL
  - Use standard R syntax to manipulate database data
- Parallel, distributed algorithms
  - Scalability and performance
  - Exposes in-database algorithms from ODM
  - Additional R-based algorithms executing and 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



OAA / Oracle R Enterprise 1.5.1

Predictive Analytics algorithms in-Database

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

Classification

- Decision Tree
- Logistic Regression
- Naïve Bayes
- Support Vector Machine
- RandomForest

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
- Expectation Maximization*

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

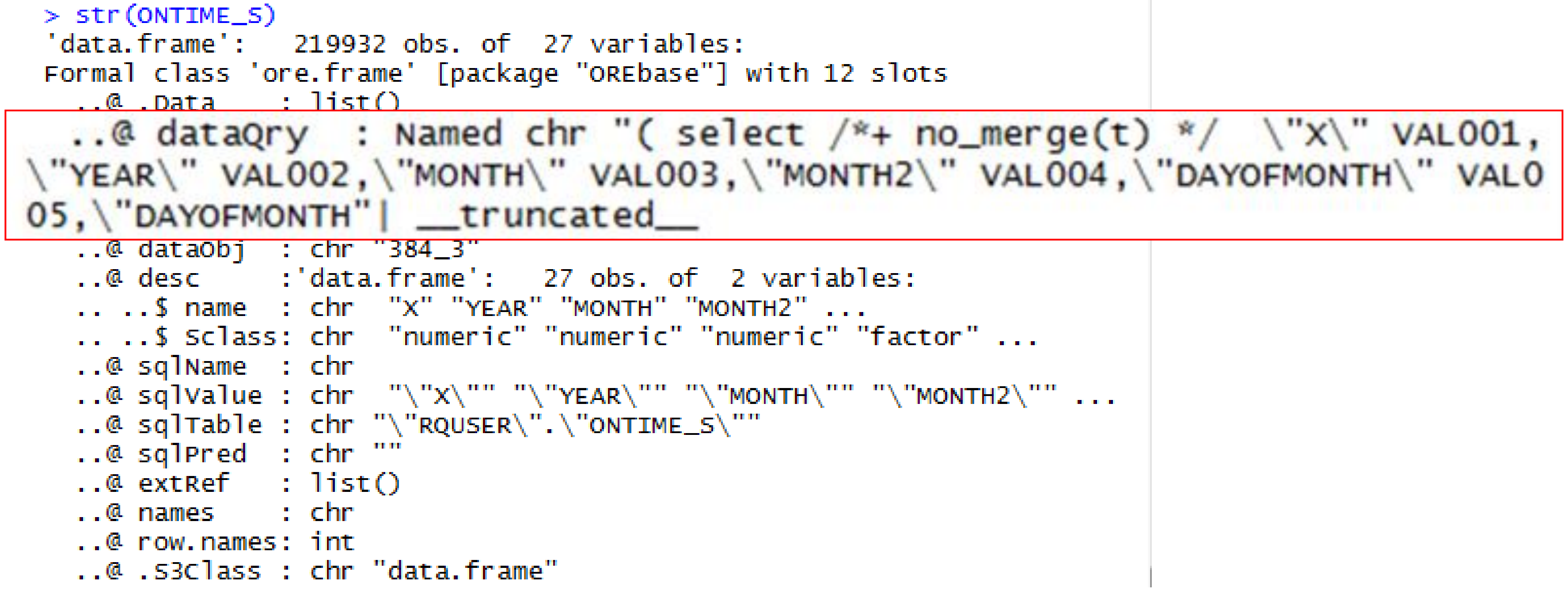
- Single Exponential Smoothing
- Double Exponential Smoothing

New in ORE 1.5.1
ODB 12.2 only

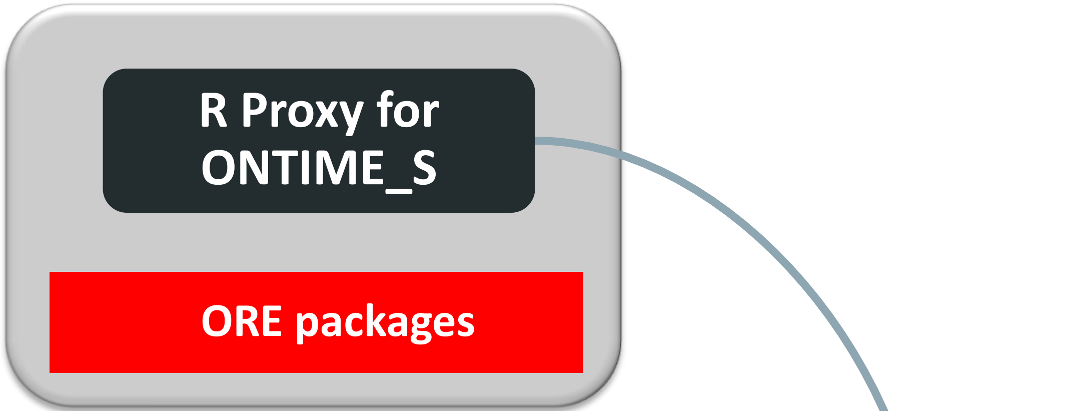
ORACLE

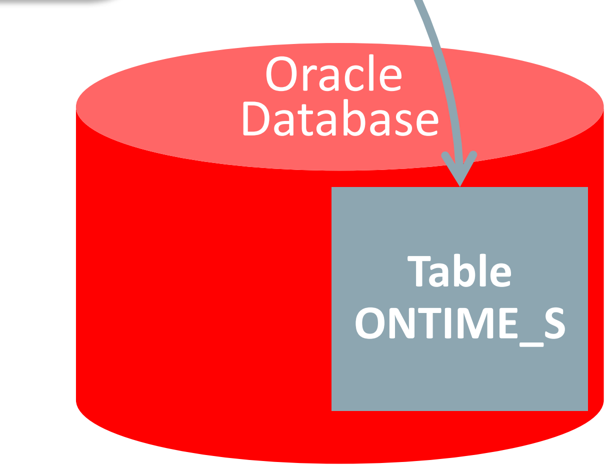
Proxy Object – ore.frame

- Inherits from data.frame
- Overloaded R functions translate functionality to SQL
- No data movement



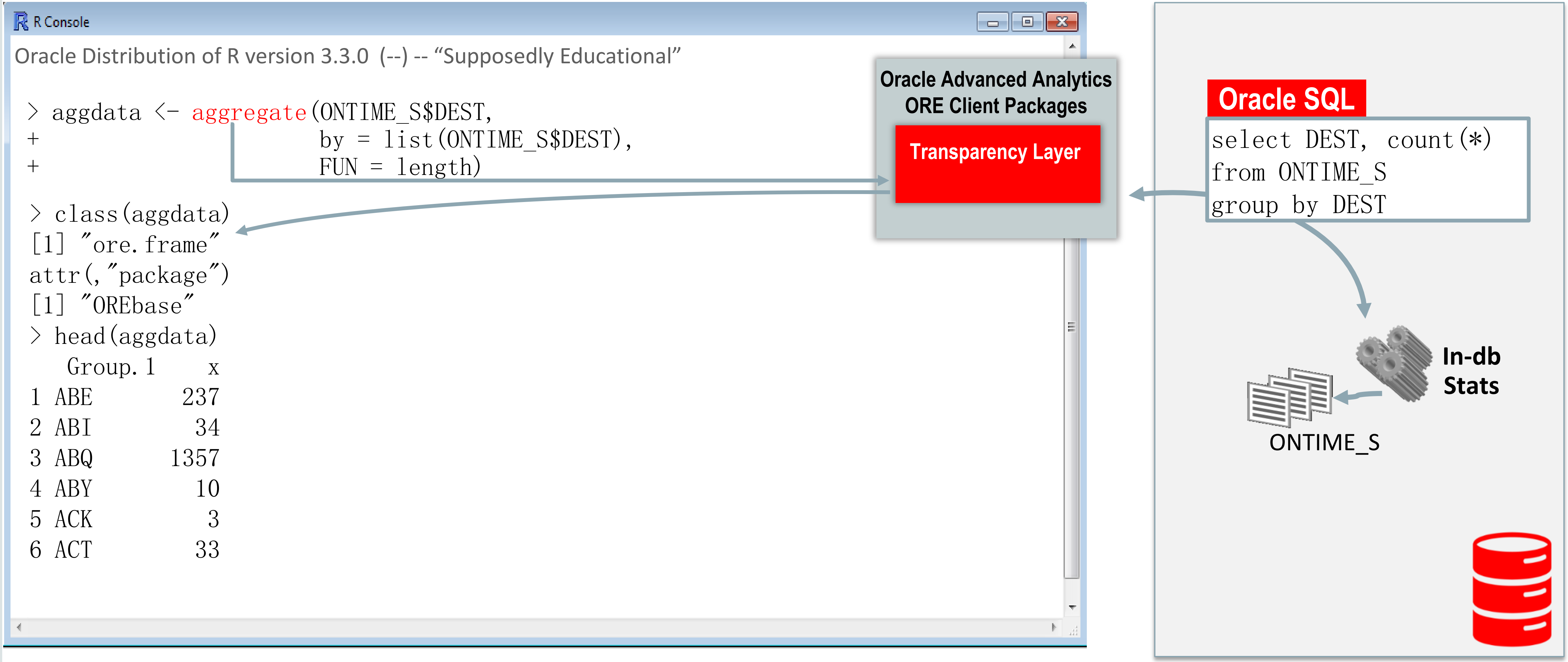
```
> str(ONTIME_S)
'data.frame':   219932 obs. of  27 variables:
Formal class 'ore.frame' [package "OREbase"] with 12 slots
  ..@ .Data     : list()
  ..@ dataQry   : Named chr "( select /*+ no_merge(t) */  \"X\" VAL001,
\"YEAR\" VAL002,\"MONTH\" VAL003,\"MONTH2\" VAL004,\"DAYOFMONTH\" VAL0
05,\"DAYOFMONTH"| __truncated__
  ..@ dataObj   : chr "384_3"
  ..@ desc      :'data.frame':   27 obs. of  2 variables:
  .. ..$ name  : chr  "X" "YEAR" "MONTH" "MONTH2" ...
  .. ..$ Sclass: chr  "numeric" "numeric" "numeric" "factor" ...
  ..@ sqlName   : chr
  ..@ sqlValue  : chr  "\"X\"" "\"YEAR\"" "\"MONTH\"" "\"MONTH2\"" ...
  ..@ sqlTable  : chr "\"RQUSER\".\"ONTIME_S\""
  ..@ sqlPred   : chr ""
  ..@ extRef    : list()
  ..@ names     : chr
  ..@ row.names: int
  ..@ .S3Class  : chr "data.frame"
```





Scalability through proxies with function overloading

In-database aggregation – no data movement

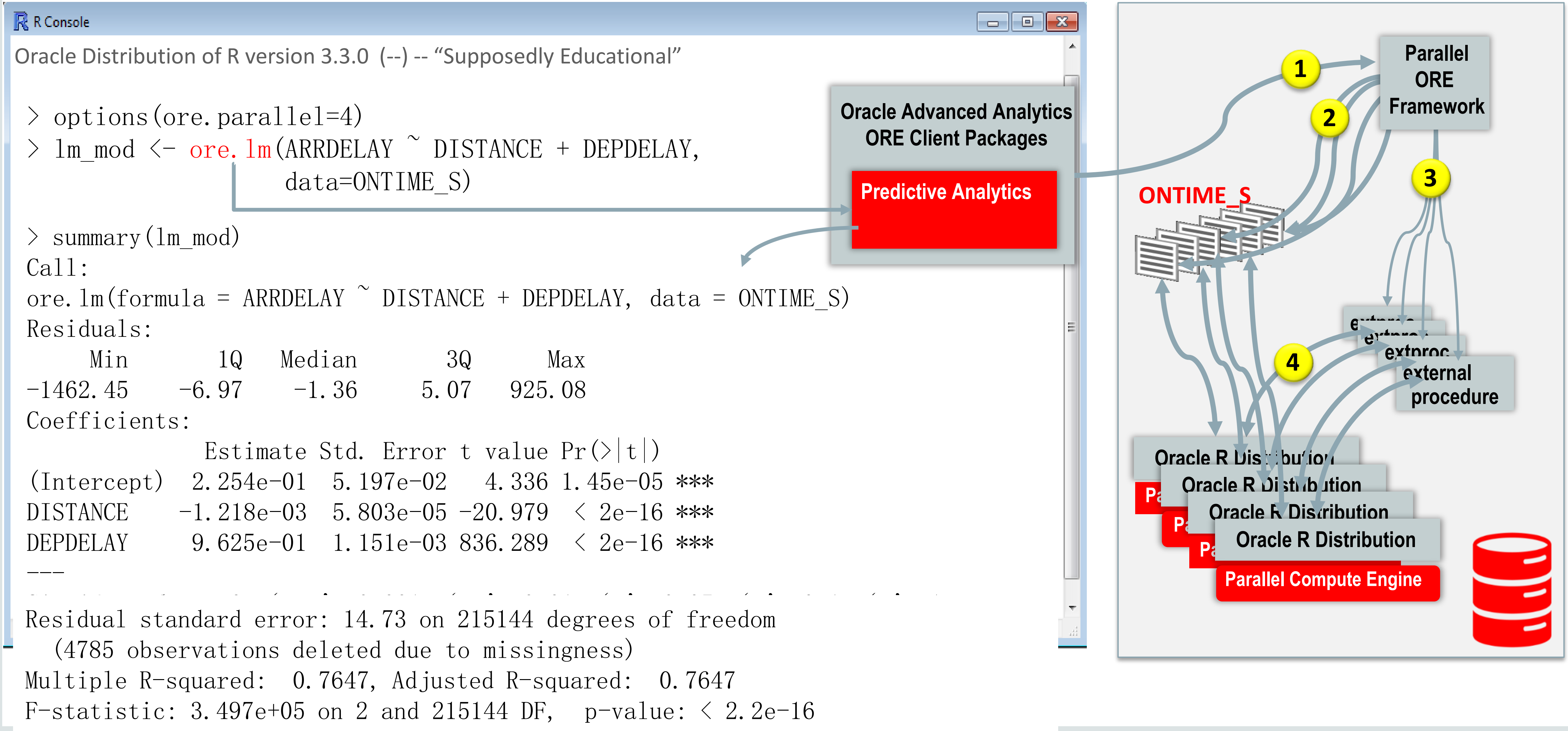


```
Oracle Distribution of R version 3.3.0  (--) -- “Supposedly Educational”

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

 > class(aggdata)
 [1] ”ore.frame”
 attr(,”package”)
 [1] ”OREbase”
 > head(aggdata)
    Group.1     x
 1  ABE       237
 2  ABI        34
 3  ABQ      1357
 4  ABY        10
 5  ACK         3
 6  ACT        33
```

```
select DEST, count(*)
from ONTIME_S
group by DEST
```

Scalable Machine Learning Algorithms

ORE parallel distributed model (e.g., Linear Regression) using embedded R engines



R Console

Oracle Distribution of R version 3.3.0 (--) -- “Supposedly Educational”

```
> options(ore.parallel=4)
> lm_mod <- ore.lm(ARRDELAY ~ DISTANCE + DEPDELAY,
                  data=ONTIME_S)

> summary(lm_mod)
Call:
ore.lm(formula = ARRDELAY ~ DISTANCE + DEPDELAY, data = ONTIME_S)
Residuals:
     Min       1Q   Median       3Q      Max
-1462.45    -6.97    -1.36     5.07   925.08
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.254e-01  5.197e-02   4.336 1.45e-05 ***
DISTANCE    -1.218e-03  5.803e-05 -20.979  < 2e-16 ***
DEPDELAY     9.625e-01  1.151e-03 836.289  < 2e-16 ***
---
Residual standard error: 14.73 on 215144 degrees of freedom
  (4785 observations deleted due to missingness)
Multiple R-squared:  0.7647, Adjusted R-squared:  0.7647
F-statistic: 3.497e+05 on 2 and 215144 DF,  p-value: < 2.2e-16
```

Linear Model Performance Comparison

- Predict “Total Revenue” of a customer based on 31 numeric variables as predictors, on 184 million records using SPARC T5-8, 4TB of RAM
- Data in an Oracle Database table

Algorithm	Threads Used*	Memory required**	Time for Data Loading***	Time for Computation	Total	Relative Performance
Open-Source R Linear Model (lm)	1	220Gb	1h3min	43min	1h46min	1x
Oracle R Enterprise lm (ore.lm)	1	-	-	42.8min	42.8min	2.47X
Oracle R Enterprise lm (ore.lm)	32	-	-	1min34s	1min34s	67.7X
Oracle R Enterprise lm (ore.lm)	64	-	-	57.97s	57.97s	110X
Oracle R Enterprise lm (ore.lm)	128	-	-	41.69s	41.69s	153X

*Open-source R lm() is single threaded

**Data moved into the R Session’s memory, since open-source lm() requires all data to be in-memory

***How long it takes to load 40Gb of raw data into the open-source R Session’s memory

IoT Use Case: Sensor Data Analysis

Massive Predictive Modeling

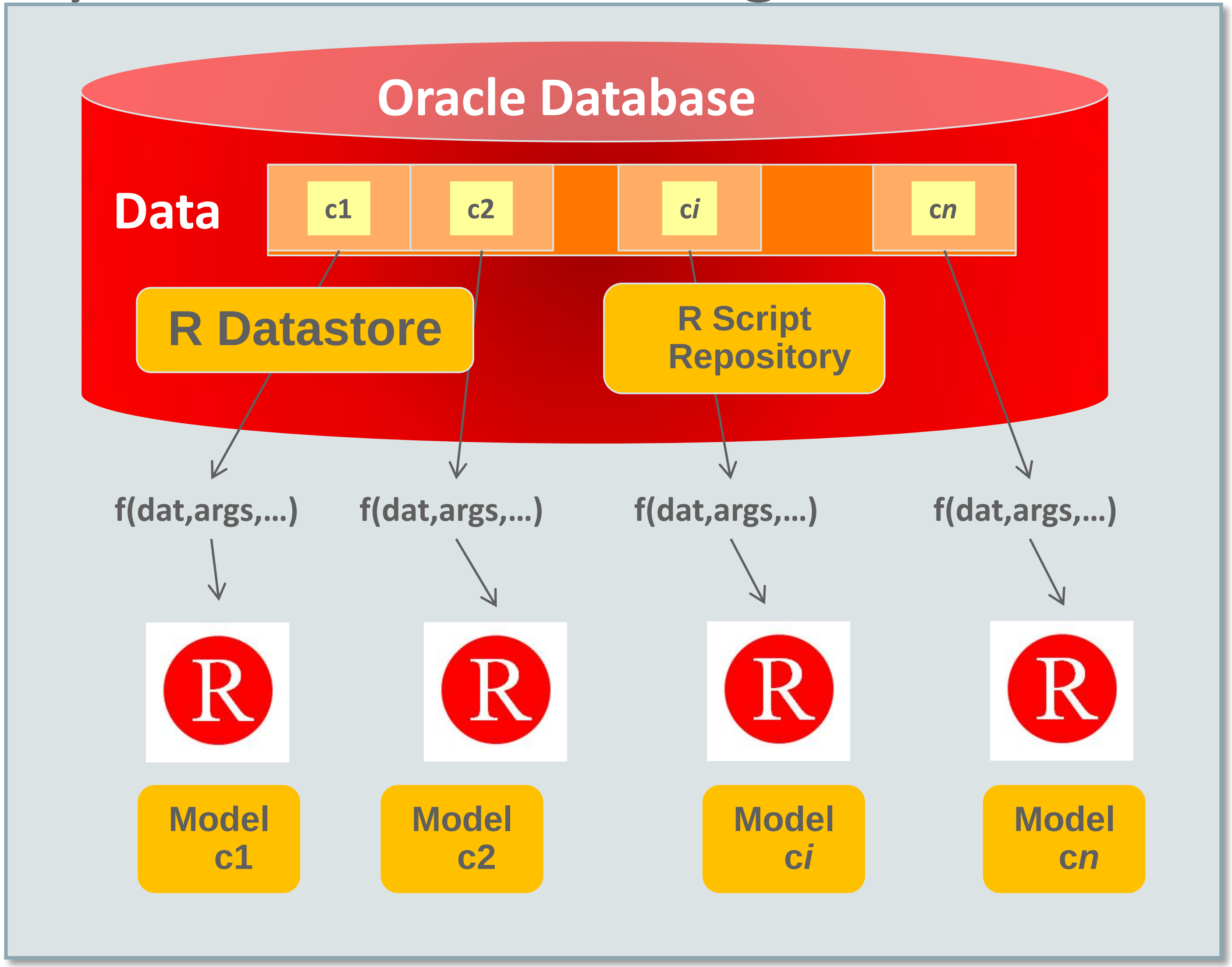
- Model each customer’s behavior and identify deviations in individual behavior and overall aggregate demand
- 200 thousand households, each with a utility “smart meter”
- 1 reading / meter / hr
- 200K x 8760 hrs / yr → 1.752B readings
- 3 years worth of data → 5.256B readings
- Each customer has 26280 readings
- If each model takes 10 seconds to build, 555.6 hrs (23.2 days)

  …with 128 DOP → 4.3 hrs

Scalable Sensor Data Analysis – Model Building

Smart meter scenario

```
f(dat,args,...) {
    R Script
    build
    model
}
```



Build models and store in database, partition on CUST_ID

14 lines

```
ore.groupApply (CUST_USAGE_DATA,
                CUST_USAGE_DATA$CUST_ID,
    function(dat, ds.name) {
      cust_id <- dat$CUST_ID[1]
      mod <- lm(Consumption ~ . -CUST_ID, dat)
      mod$effects <- mod$residuals <- mod$fitted.values <- NULL
      name <- paste("mod", cust_id,sep="")
      assign(name, mod)
      ds.name1 <- paste(ds.name,".",cust_id,sep="")
      ore.save(list=paste("mod",cust_id,sep=""), name=ds.name1, overwrite=TRUE)
      TRUE
    },
    ds.name="myDatastore", ore.connect=TRUE, parallel=TRUE
  )
```

Production Deployment of R through SQL

- Load R function into Oracle Database from R or SQL
- From SQL
  - Return images as PNG BLOB column
  - Return data.frame content as database table
  - Return XML with image and data.frame content
- Invoke same function from R

Have fun and
raise your hand if you need help

Learn More about
Oracle's Advanced Analytics R Technologies…

http://oracle.com/goto/R

R Technologies from Oracle

Bringing the Power of R to the Enterprise

ORACLE®