Taking R to New Heights for Scalability and Performance

Mark Hornick
Director, Advanced Analytics and Machine Learning
mark.hornick@oracle.com
@MarkHornick
blogs.oracle.com/R

January 31,2017

Safe Harbor Statement

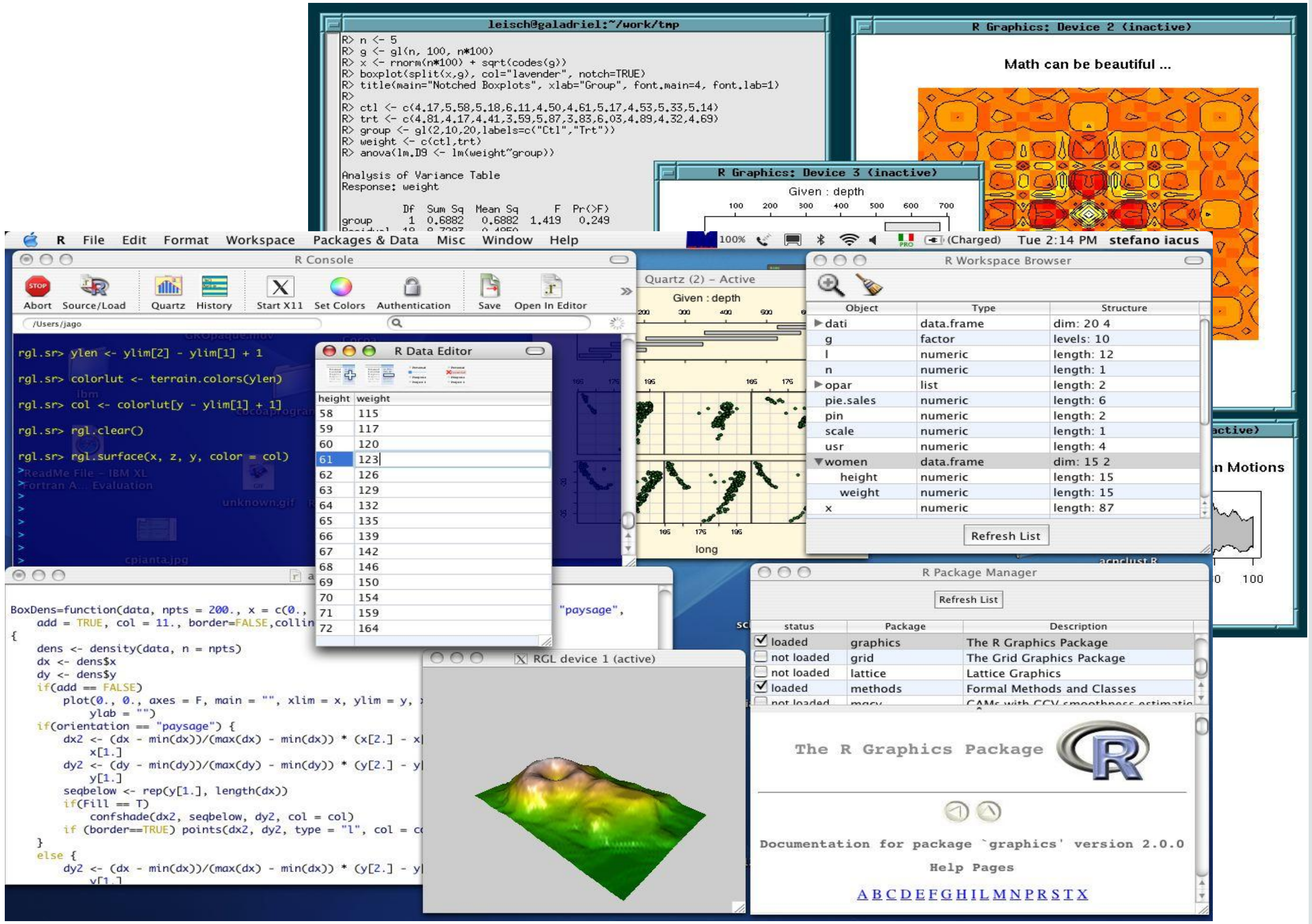
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development,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

Why statisticians | data analysts | data scientists use R

R is a statistics language similar to Base SAS or SPSS statistics

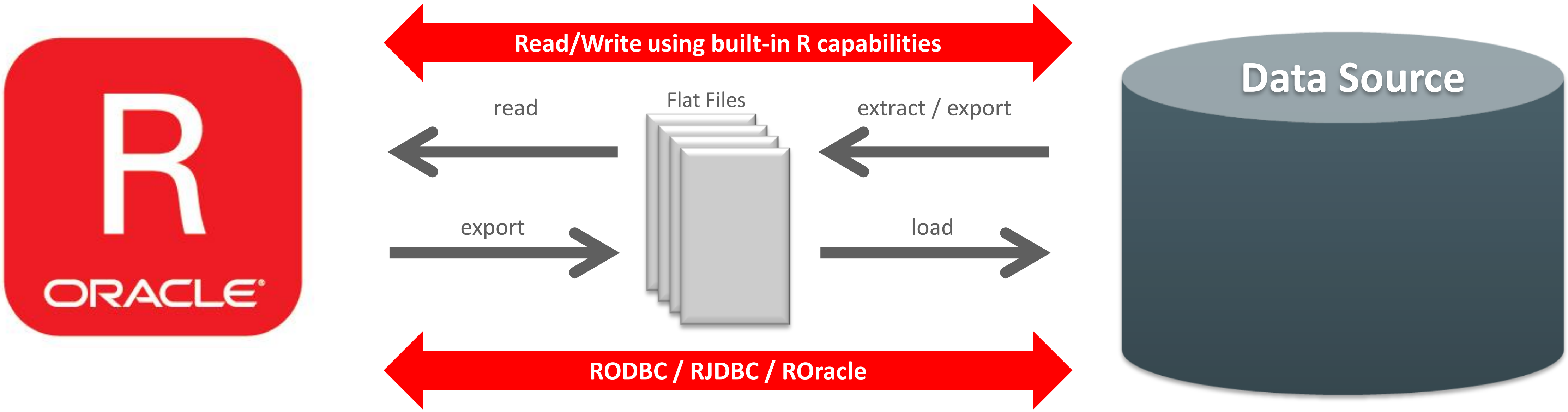
- Powerful
- Extensible
- Graphical
- Extensive statistics
- Ease of installation and use
- Rich ecosystem
  - ~10K open source packages
  - Millions of users worldwide
- ***Free***

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



ORACLE

Traditional R and Data Source Interaction



Deployment
R script
cron job

- Access latency
- Paradigm shift: R → *Data Access Language* → R
- Memory limitation – data size, call-by-value
- Single threaded
- Ad hoc production deployment
- Issues for backup, recovery, security

ORACLE

How to take R to new heights for scalability and performance?

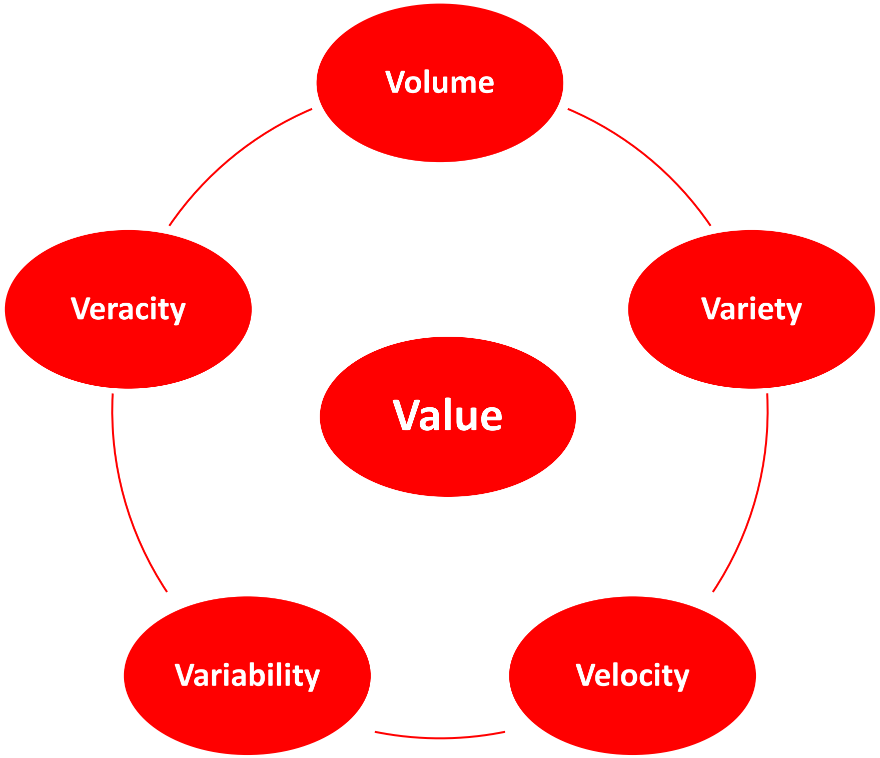
i.e., to work on Big Data

big da·ta

noun COMPUTING

extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
"much IT investment is going towards managing and maintaining big data"

https://www.google.com/search?q=big+data&ie=utf-8&oe=utf-8

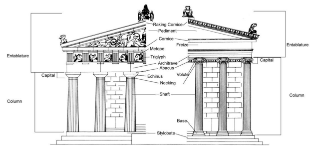


Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.[2] "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[3]

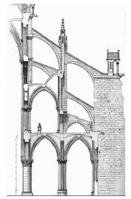
https://en.wikipedia.org/wiki/Big_data

ORACLE

Capabilities that take R to new heights…

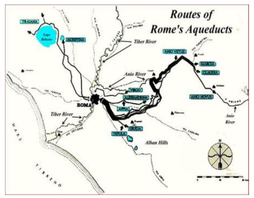


Transparent Data Access, Analysis, and Exploration



Production Deployment

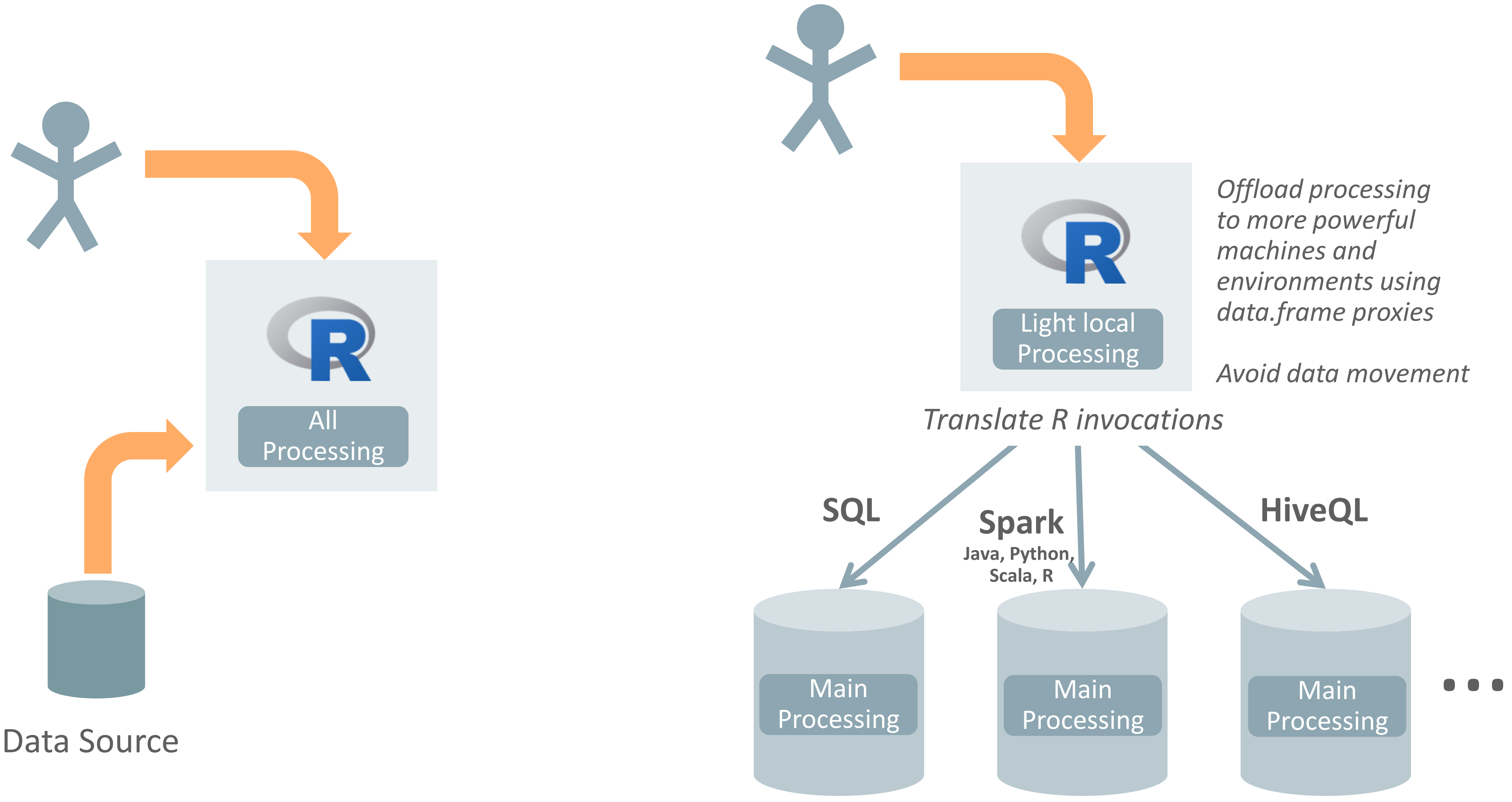
Scalable Machine Learning



Data & Task Parallelism

ORACLE

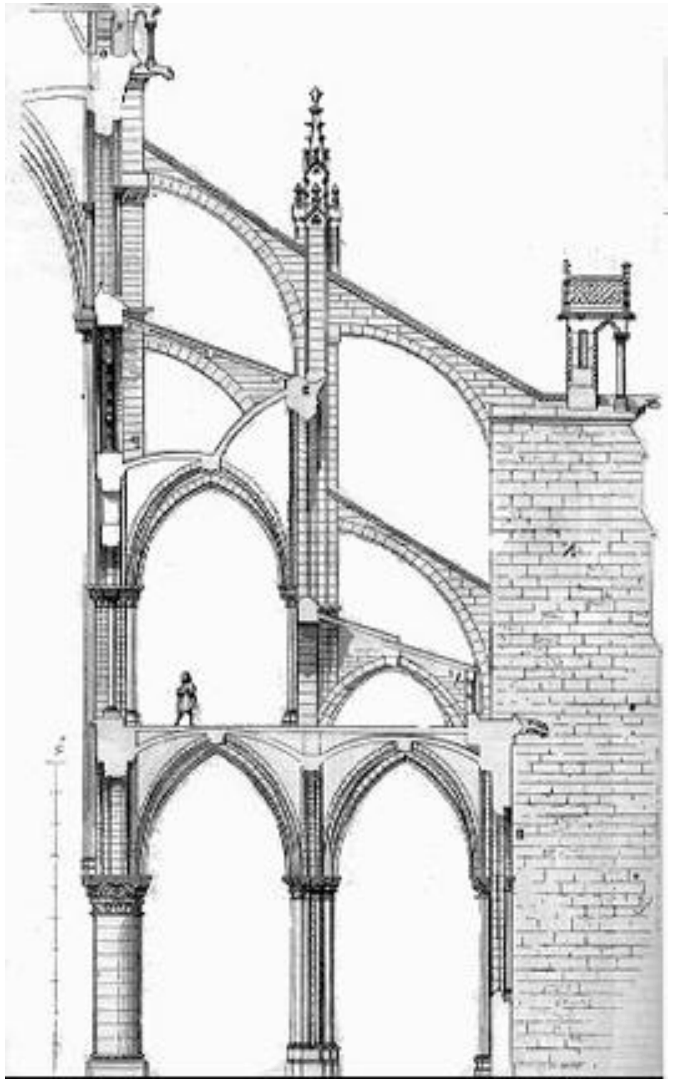
Transparent data access, analysis, and exploration



Transparent data access and manipulation

- Maintain language features and interface
- Transparently translate R to language of powerful data processing engines
- Reference data to eliminate data movement

- Analyze all of your data



Proxy objects for Big Data

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

data.frame

Inherits from

Proxy
data.frame

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> str(IRIS)
'data.frame':   150 obs. of  5 variables:
Formal class 'ore.frame' [package "OREbase"] with 12 slots
  ..@ .Data     : list()
  ..@ dataQry   : Named chr "( select /*+ no_merge(t) */  \"Sepal.Length\" VAL001,\"Sepal.Wid
```

```
"( select /*+ no_merge(t) */  \"Sepal.Length\" VAL001,\"Sepal.Width\" VAL
002,\"Petal.Length\" VAL003,\"Petal.Width\" VAL004,\"Species\" VAL005 fro
m \"RQUSER\".\"IRIS\" t  )"
```

```
  .. ..$ S3Class: chr  "numeric" "numeric" "numeric" "numeric" ...
  ..@ sqlName   : chr
  ..@ sqlValue  : chr  "\"Sepal.Length\"" "\"Sepal.Width\"" "\"Petal.Length\"" "\"Petal.Width
\"" ...
  ..@ sqlTable  : chr "\"RQUSER\".\"IRIS\""
  ..@ sqlPred   : chr ""
  ..@ extRef    : list()
  ..@ names     : chr
  ..@ row.names: int
  ..@ .S3Class  : chr "data.frame"
```

ORACLE

Transparency Examples

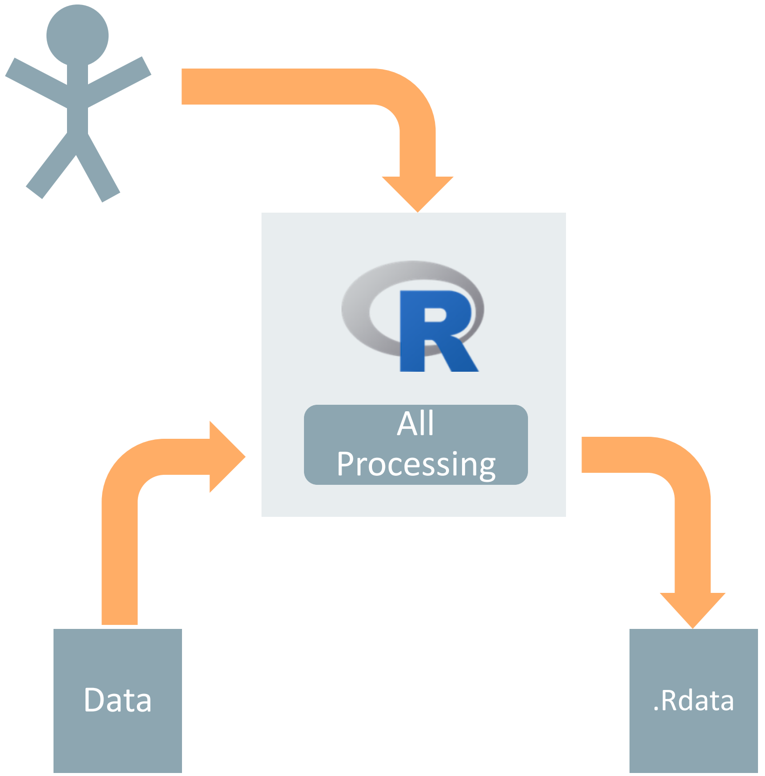
```
library(ORE)
ore.connect("rquser", "orcl",
  "localhost", "rquser", all=TRUE)
ore.ls()

df <- with(ONTIME_S,
 ONTIME_S[DEST=="SFO"|DEST=="BOS",1:21])

df$LRGDELAY <-
  ifelse(df$ARRDELAY > 20,1,0)
head(df)
summary(df)
```

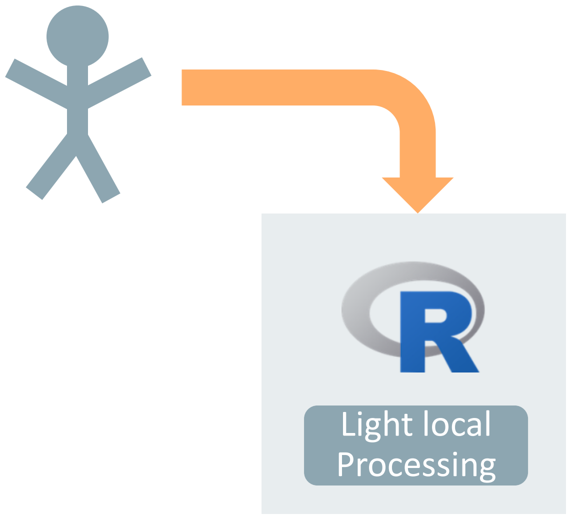
```
hist(MY_TABLE$ARRDELAY,breaks=100)

merge (TEST_DF1, TEST_DF2,
       by.x="x1", by.y="x2")


# with OREdplyr in ORE 1.5.1...

select(FLIGHTS, year, month, dep_delay)
rename(FLIGHTS, tail_num = tailnum)
filter(FLIGHTS, month == 1, day == 1)
arrange(FLIGHTS, year, month, day)
mutate(FLIGHTS, speed=air_time/distance)
```

ore.frame Proxy Object

ORACLE

Scalable Machine Learning

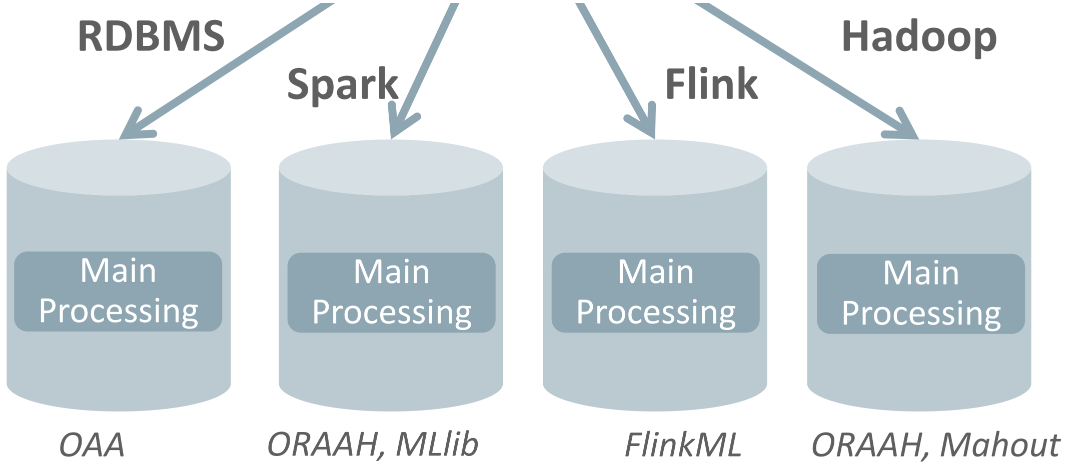




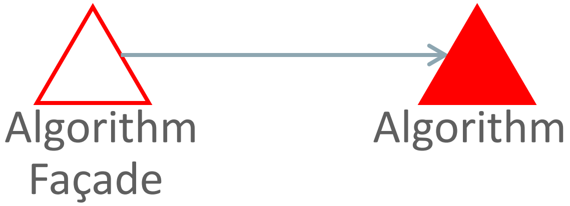
Offload model building to parallel software and powerful machines and environments

Enable new technologies as they arise

Build models using powerful hardware and parallel algorithms…



Scalable Machine Learning



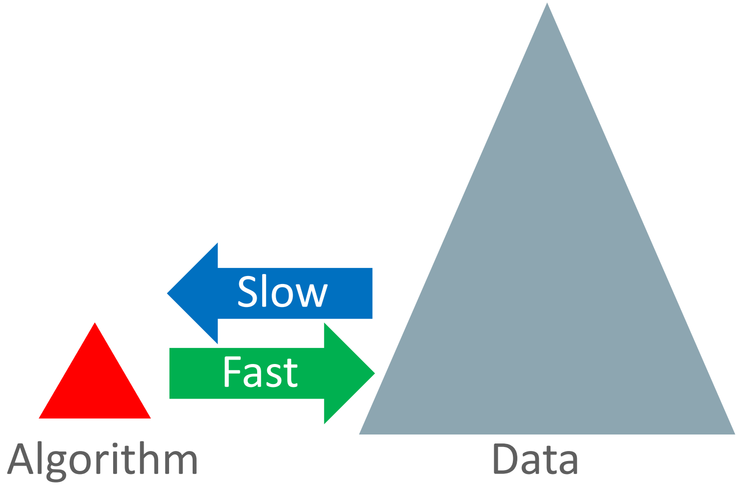
- Maintain R machine learning interface
  - Easy to specify formula – minimal lines of code
  - Include transformations, interaction terms, etc.

Target **Predictors**

log(ARRDELAY) ~ DISTANCE + ORIGIN + DEST +
as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) +
as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)

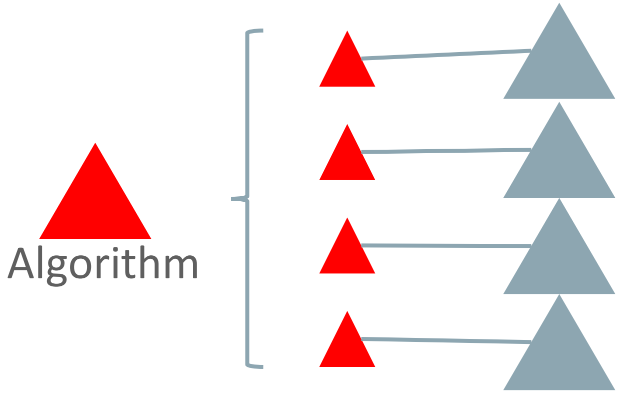
Scalable Machine Learning

- Maintain R machine learning interface
  - Easy to specify formula – minimal lines of code
  - Include transformations, interaction terms, etc.
- Bring the algorithm to the data
  - Eliminate or minimize data movement
  - Leverage proxy objects to reference data



Scalable Machine Learning

- Maintain R machine learning interface
  - Easy to specify formula – minimal lines of code
  - Include transformations, interaction terms, etc.
- Bring the algorithm to the data
  - Eliminate or minimize data movement
  - Leverage proxy objects to reference data
- Parallel, distributed algorithm implementations
  - Oracle-proprietary parallel, distributed algorithms
  - Leverage other open source packages and toolkits
    e.g., Apache Spark Mllib, Apache FlinkML



ORACLE

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

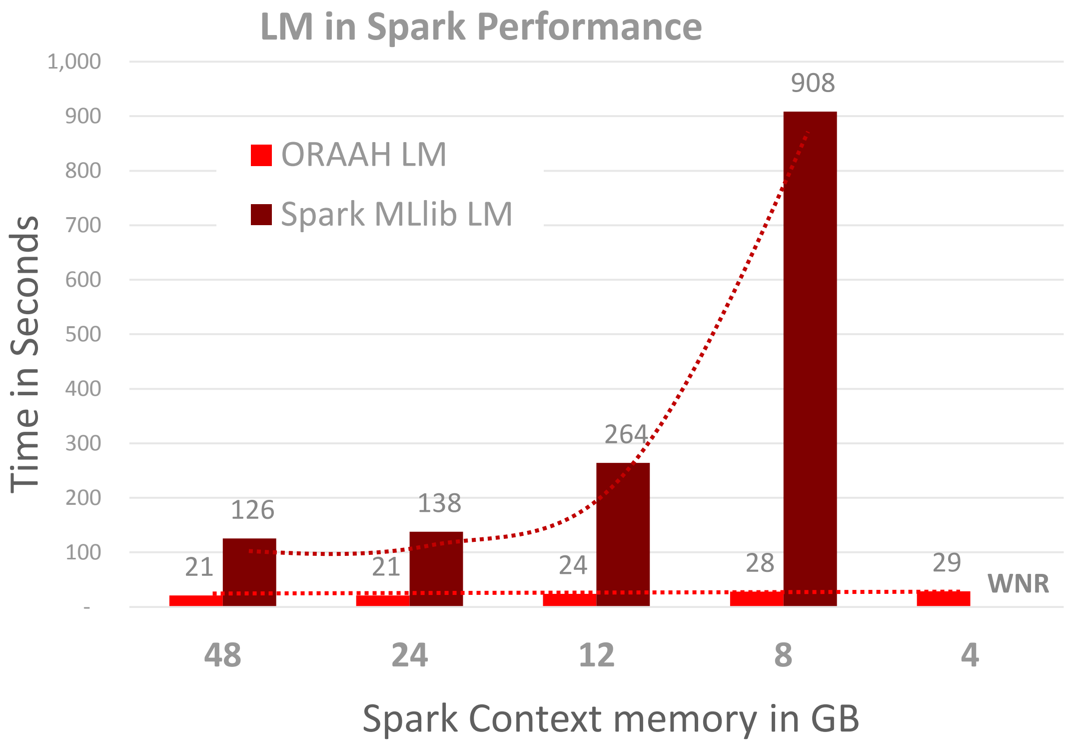
ORACLE

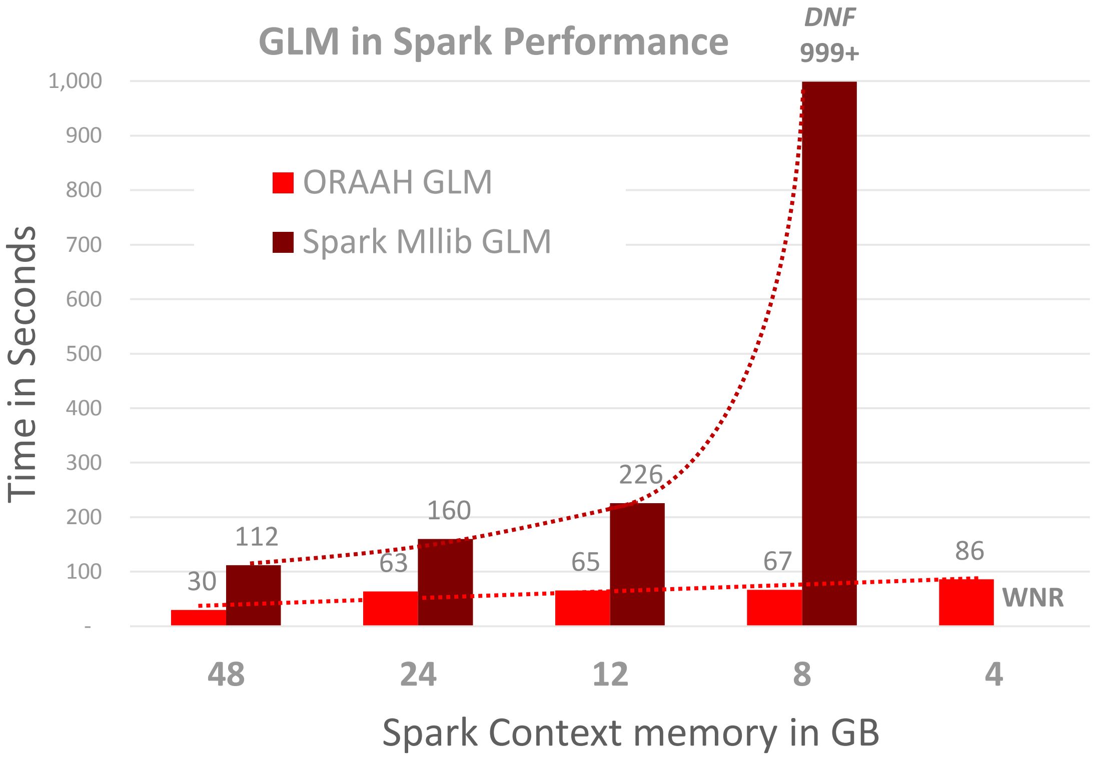
Not all parallel implementations are the same

Comparing performance with varying Spark memory footprints

Benchmark on single X5-2 Node with 74 threads and 256 GB of Total RAM, Spark 1.6.0 on CDH 5.8.0

Input Data is 15GB ”Ontime” airline dataset with 123mi records, predicting 8,926 total coefficients

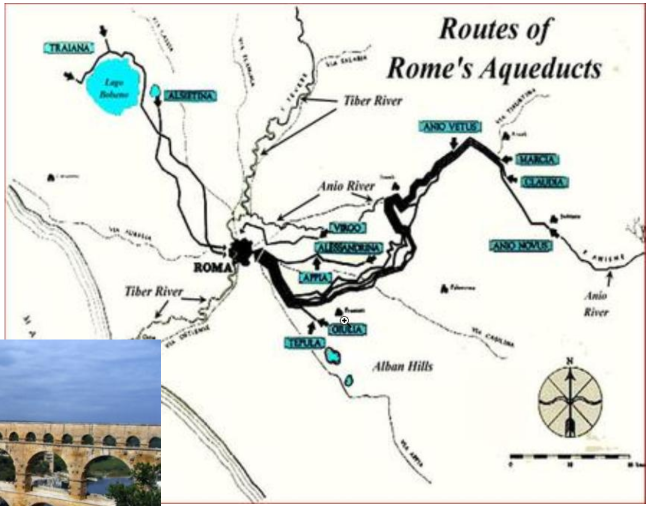


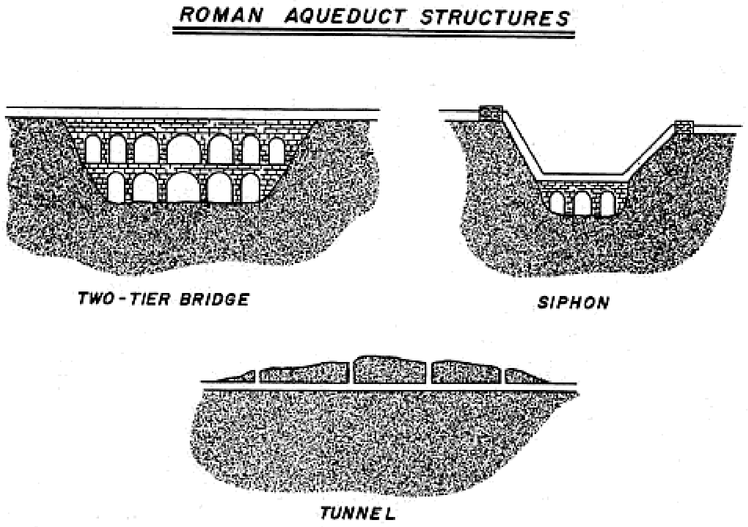


LM formula used ARRDELAY ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)
GLM formula used CANCELLED ~ DISTANCE + ORIGIN + DEST + as.factor(MONTH) + as.factor(YEAR) + as.factor(DAYOFMONTH) + as.factor(DAYOFWEEK) + as.factor(FLIGHTNUM)

Data and Task Parallel Execution

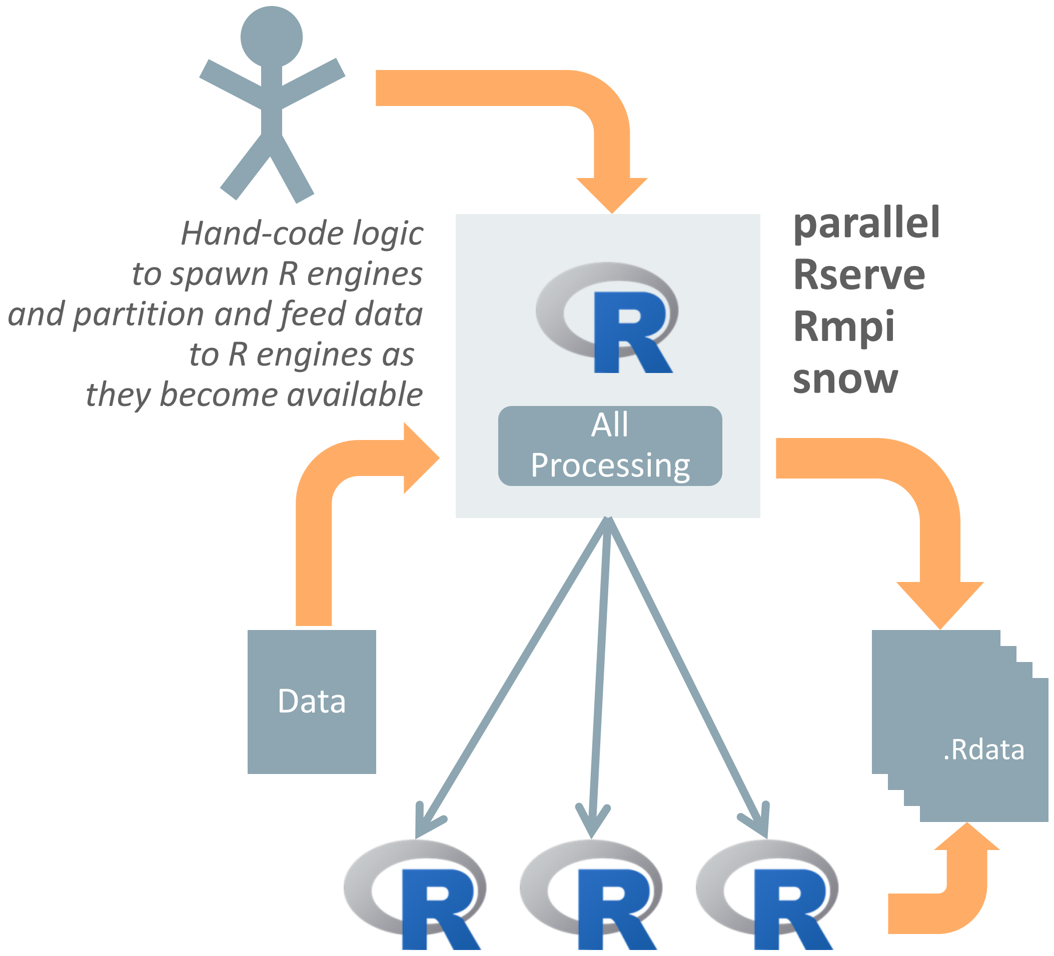
- Easily specify parallelism and data partitioning
  - Simplified API – *all-in-one*
  - Build and score with millions of models
- Automated management of parallel R engines
  - Insulation from hardware details
  - Limit resources as appropriate
  - Startup and shutdown automatically
- Automated loading of data into parallel R engines
- Leverage CRAN packages

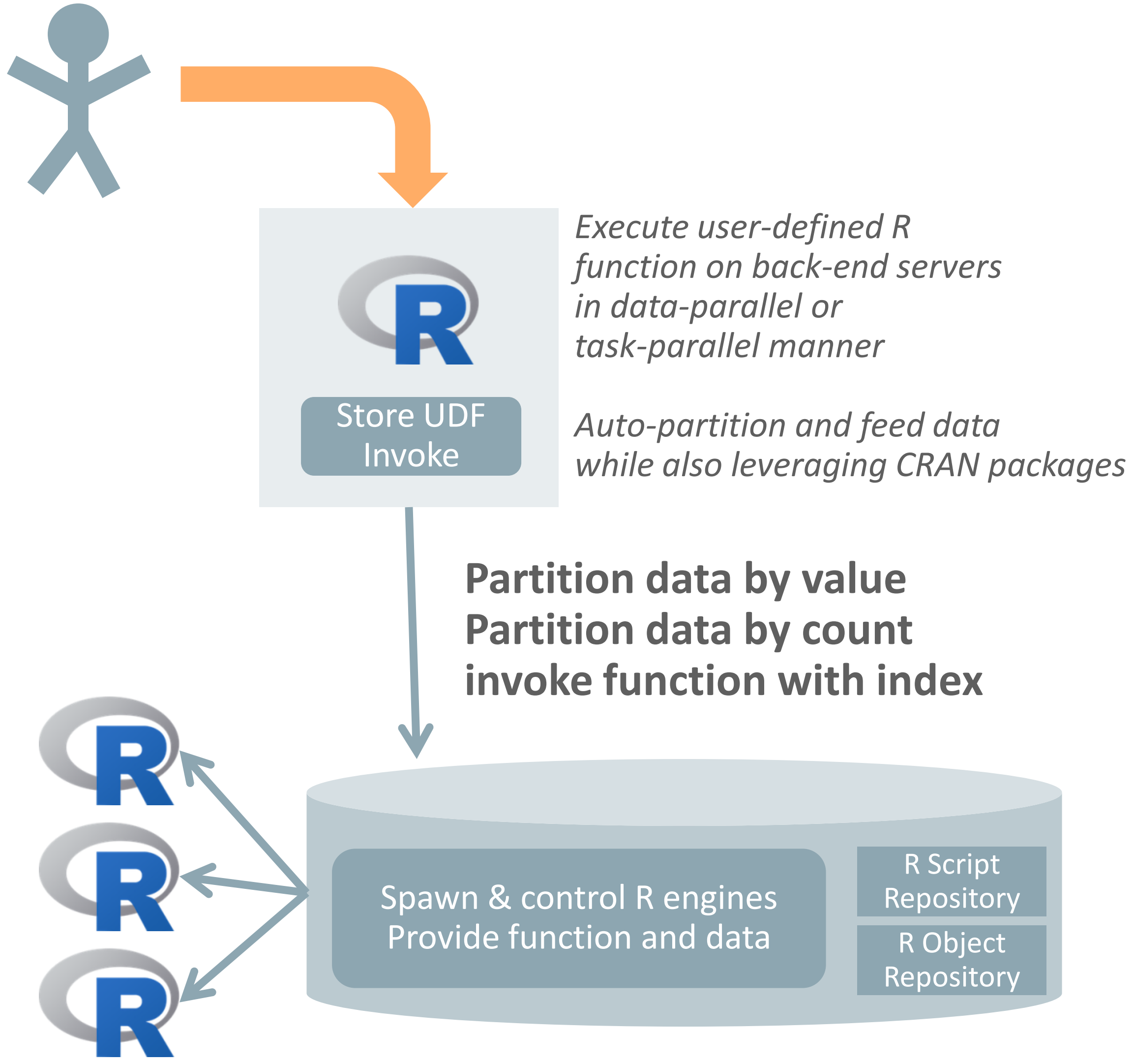




ORACLE

Data and Task Parallelism





Example API

- Supply data
- Specify function
- Use CRAN packages
- Store and load R objects
- Pass Arguments
- Specify parallelism
- Get/use results
  - R objects
  - structured data
  - Images
  - etc.

No parallelism

```
library(e1071)
mod <- ore.tableApply(IRIS_TABLE,
                      function(dat,datastore) {
                        library(e1071)
                        dat$Species <- as.factor(dat$Species)
                        mod<-naiveBayes(Species ~ ., dat)
                        ore.save(mod,name=datastore)
                      },
                      datastore="NB_Model-1")
```

Data parallel by *chunk*

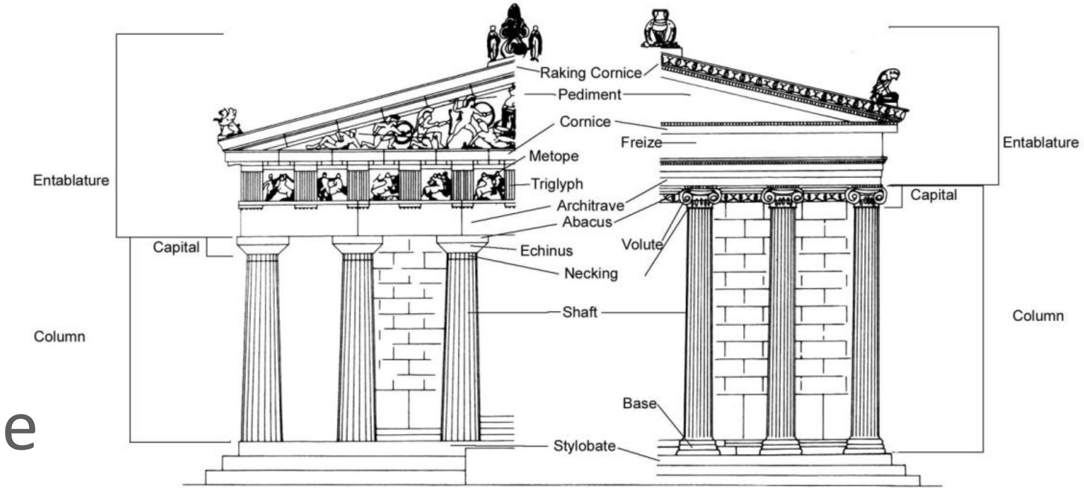
```
scoreNBmodel <- function(dat, datastore) {
     library(e1071)
     ore.load(datastore)
     dat$PRED <- predict(mod, newdata = dat)
     dat
  }

IRIS_PRED      <- IRIS_TABLE[1,]
IRIS_PRED$PRED <- "A"

res <- ore.rowApply(IRIS_TABLE, scoreNBmodel, datastore = "NB_Model-1",
                    parallel=4, FUN.VALUE=IRIS_PRED, rows=10)
```

Data parallel by *partition*

```
DAT <- ONTIME_S[ONTIME_S$DEST %in%
                c("BOS","SFO","LAX","ORD","ATL","PHX","DEN"),]

modList <- ore.groupApply(
    X=DAT, INDEX=DAT$DEST, parallel=3,
    function(dat) lm(ARRDELAY ~ DISTANCE + DEPDELAY, dat))
```

ORACLE

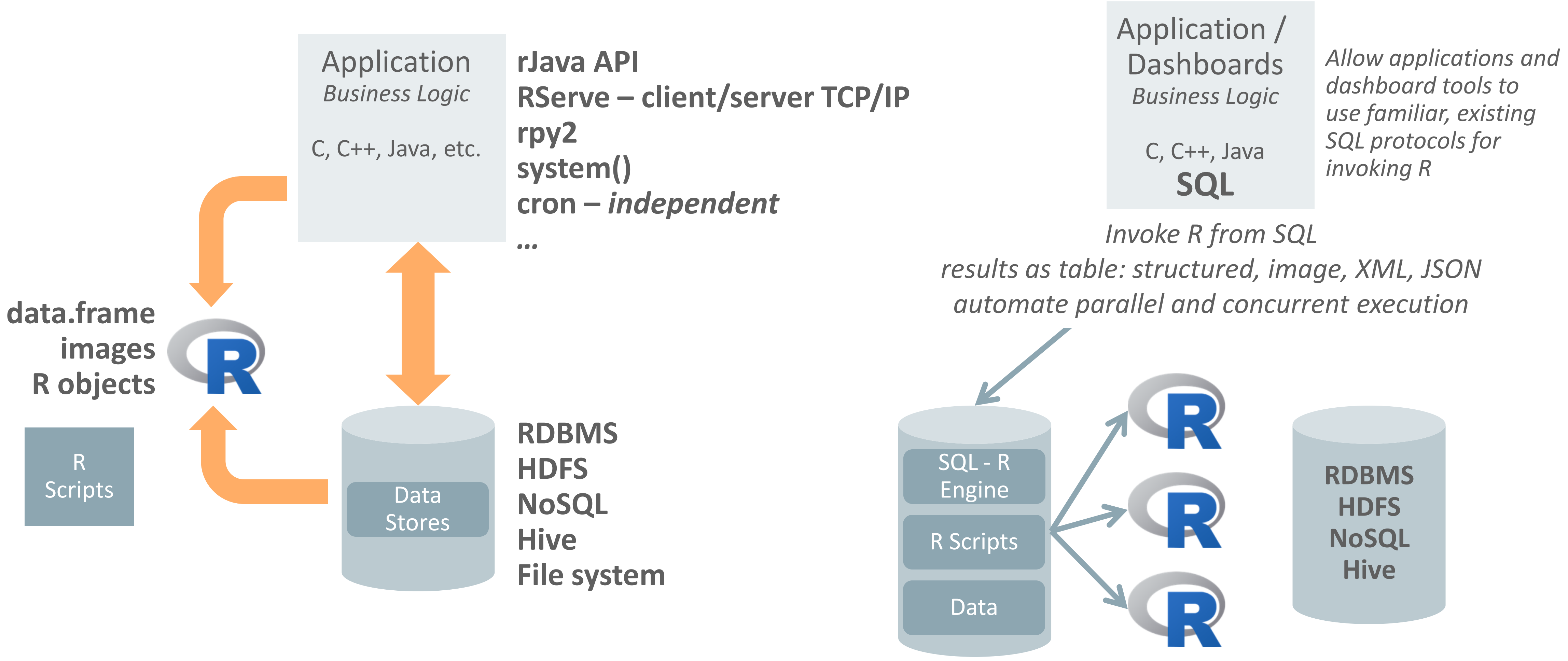
Deployment



- Avoid costly recoding or translating R code
- Invoke R easily from non-R environments
- Map data structures and types naturally
- Seamlessly return data.frames, images, XML, JSON in local environment data structures

ORACLE

Deployment


Application
Business Logic
C, C++, Java, etc.
rJava API
RServe – client/server TCP/IP
rpy2
system()
cron – independent
…
data.frame
images
R objects
R
R
Scripts
Data
Stores
RDBMS
HDFS
NoSQL
Hive
File system
Application /
Dashboards
Business Logic
C, C++, Java
SQL
Allow applications and dashboard tools to use familiar, existing SQL protocols for invoking R
Invoke R from SQL
results as table: structured, image, XML, JSON
automate parallel and concurrent execution
SQL - R
Engine
R Scripts
Data
R
R
R
RDBMS
HDFS
NoSQL
Hive

Deploy R using SQL

- Store named R function in Script Repository from R or SQL
- Return values
  - Images as PNG BLOB column
  - data.frame content as database table
  - XML with data.frame and image
- Benefits
  - Fewer moving parts
  - IPC data transfer speeds at backend
  - Invoke same function from R or SQL
  - Security
  - Integrated backup and recovery

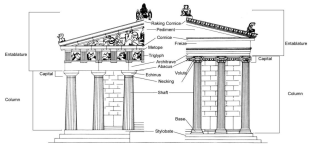
```
begin
  sys.rqScriptDrop('RandomRedDots');
  sys.rqScriptCreate('RandomRedDots',
 'function(){
            id <- 1:10
            plot( 1:100, rnorm(100), pch = 21,
                  bg = "red", cex = 2, main="Random Red Dots"
)
            data.frame(id=id, val=id / 100)
            }');
end;
```

```
select    ID, IMAGE
from      table(rqEval( NULL,'PNG','RandomRedDots'));


select    id, val
from      table(rqEval( NULL,'select 1 id, 1 val from dual',
                 'RandomRedDots'));


-- Return structured and image content within XML string
select    *
from      table(rqEval(NULL, 'XML', 'RandomRedDots'));

-- In R, invoke same function by name

 ore.doEval(FUN.NAME='RandomRedDots')
```

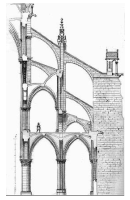
ORACLE

Architectural Elements: Enabling R for Big Data

- Leverage powerful back-ends for the heavy lifting…transparently
- Leverage new, more powerful back-ends more easily as they appear
- Enable parallelism quickly and easily for big data processing
- Immediately leverage data scientist R scripts and results in production environments

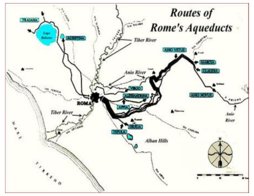


Transparent Data Access, Analysis, and Exploration



Production Deployment

Scalable Machine Learning

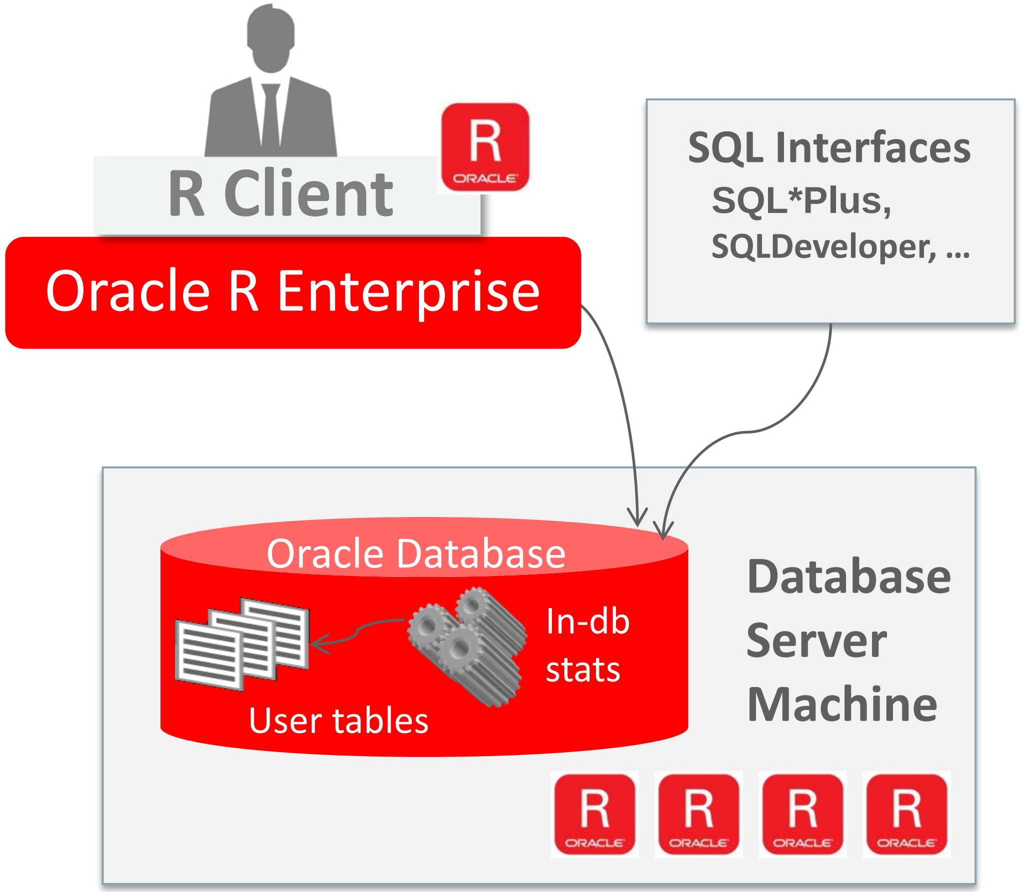


Data & Task Parallelism

Oracle R Enterprise

Part of Oracle Advanced Analytics option to Oracle Database

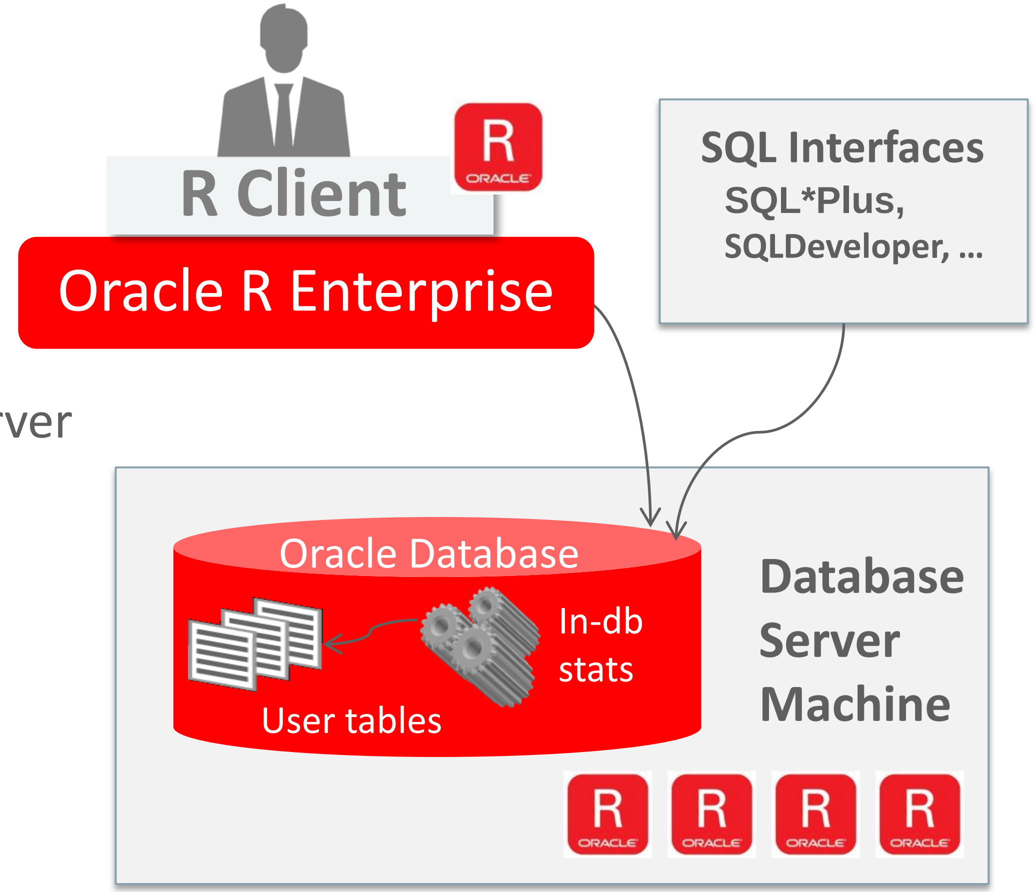
- Use Oracle Database as HPC environment
- Use in-database parallel and distributed machine learning algorithms
- Manage R scripts and R objects in Oracle Database
- Integrate R results into applications and dashboards via SQL



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

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

Market Basket Analysis
- Apriori – Association Rules

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

Attribute Importance
- Minimum Description Length
- Expectation Maximization*

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

Anomaly Detection
- 1 Class Support Vector Machine

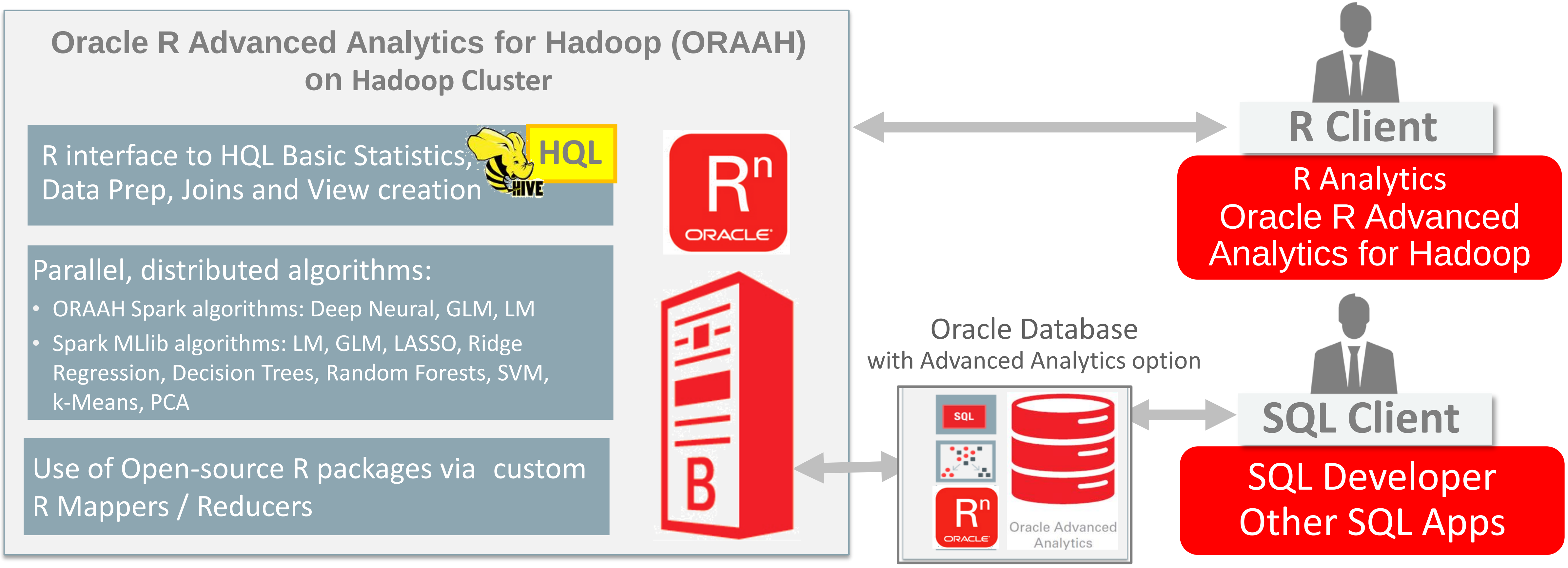
Time Series
- Single Exponential Smoothing
- Double Exponential Smoothing

New in ORE 1.5.1
ODB 12.2 only

ORACLE

Oracle R Advanced Analytics for Hadoop

Using Hadoop/Hive/Spark Integration, plus R Engine and Open-Source R Packages



ORACLE

Oracle R Advanced Analytics for Hadoop 2.7.0
Predictive Analytics algorithms

Classification

GLM ORAAH	hadoop MapReduce
Logistic Regression ORAAH	Spark
Logistic Regression	Spark MLlib
Random Forests	Spark MLlib
Decision Trees	Spark MLlib
Support Vector Machines	Spark MLlib

Clustering

Hierarchical k-Means	hadoop MapReduce
Hierarchical k-Means	Spark MLlib
Gaussian Mixture Models	Spark MLlib

Regression

MLP Neural Networks ORAAH	Spark
LASSO	Spark MLlib
Ridge Regression	Spark MLlib
Support Vector Machines	Spark MLlib
Random Forest	Spark MLlib
Linear Regression	Spark MLlib

Basic Statistics

Correlation/Covariance	hadoop MapReduce

Feature Extraction

Non-negative Matrix Factorization	hadoop MapReduce
Collaborative Filtering (LMF)	hadoop MapReduce
Singular Value Decomposition	Spark MLlib

Attribute Importance

Principal Components Analysis	hadoop MapReduce
Principal Components Analysis	Spark MLlib

ORACLE

Cloud-Based Machine Learning

- Oracle Advanced Analytics option including Oracle Data Mining and Oracle R Enterprise on:
  - Oracle Exadata Cloud Service
  - Oracle Database Cloud Service: Included in High Performance and Extreme Performance services
- Oracle R Advanced Analytics for Hadoop
  - Included in the Oracle Big Data Cloud Service

ORACLE

Demonstration of ORE

Join us for the Oracle R Enterprise Hands-on Lab Wednesday @ 9:00

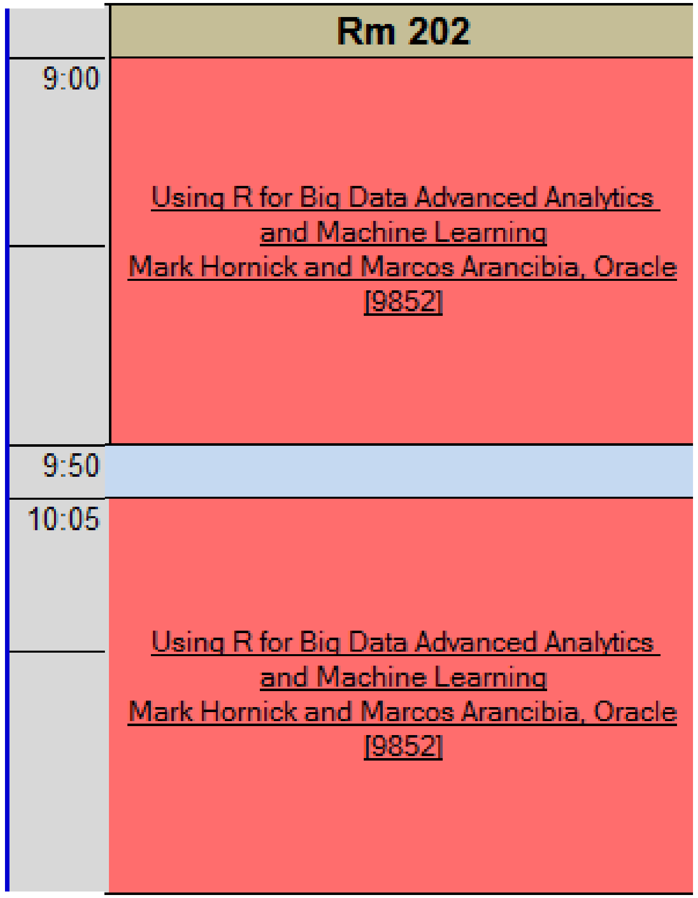
Using R for Big Data Advanced Analytics and Machine Learning

→ data exploration / attribute importance

→ clustering

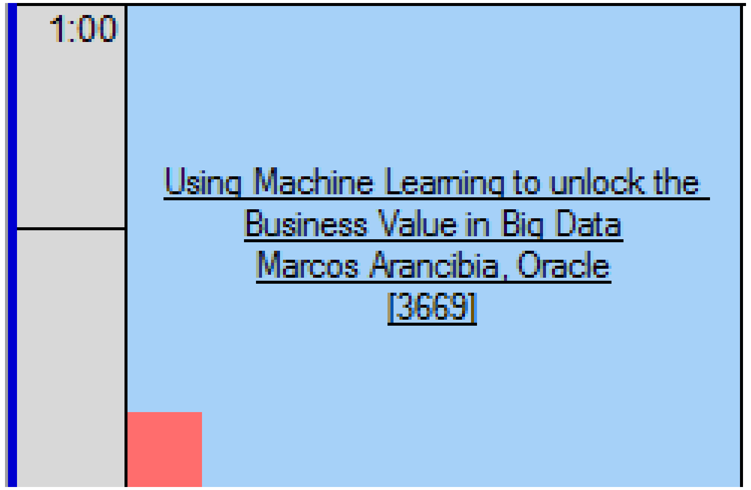
→ regression

→ OREdplyr, and more



ORACLE

Join us for Big Data with ORAAH Wednesday @ 1:00



Using Machine Learning to unlock the Business Value in Big Data

ORACLE

Join us for new technology intro Thursday @ 9:50

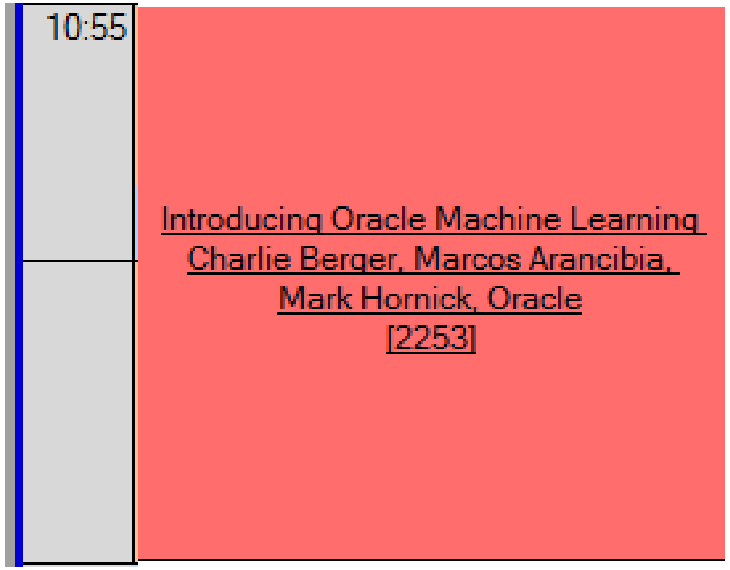
Combining Graph and Machine Learning Technologies using R

	Room 103 Spatial Summit
9:50	Combining Graph and Machine Learning Technologies using R Hassan Chafi and Mark Hornick, Oracle [1587]

ORACLE

Join us for new technology intro Thursday @ 10:55

Introducing Oracle Machine Learning
→ new notebook technology from Oracle



ORACLE

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