

# Oracle Big Data Science IOUG Collaborate 16

Session 4762

Tim and Dan VlamiS

Tuesday, April 12, 2016

# VlamiS Software Solutions

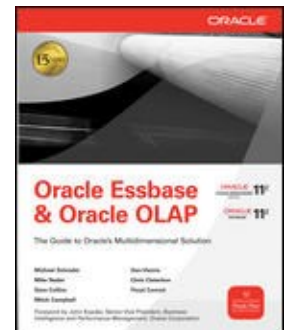
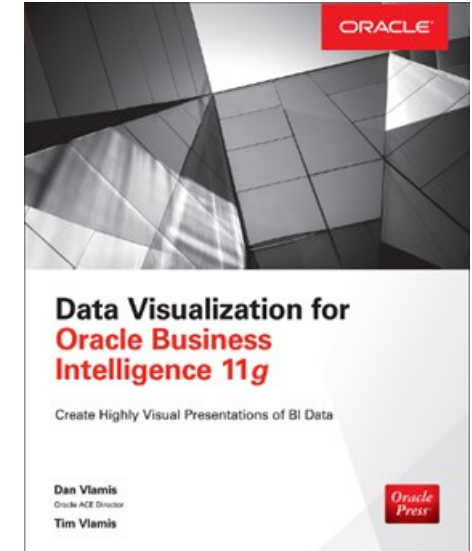
- VlamiS Software founded in 1992 in Kansas City, Missouri
- Developed 200+ Oracle BI and analytics systems
- Specializes in Oracle-based:
  - Enterprise Business Intelligence & Analytics
  - Analytic Warehousing
  - Data Mining and Predictive Analytics
  - Data Visualization
- Multiple Oracle ACEs, consultants average 15+ years
- [www.vlamiS.com](http://www.vlamiS.com) (blog, papers, newsletters, services)
- Co-authors of book “Data Visualization for OBI 11g”
- Co-author of book “Oracle Essbase & Oracle OLAP”
- Oracle University Partner
- Oracle Gold Partner

 EDUCATION RESELLER

 APPROVED  
EDUCATION CENTER

 Gold  
Partner

Specialized  
Oracle Business Intelligence  
Foundation Suite 11g







# Dan and Tim Vlami

## Dan Vlami – President

- Founded Vlami Software Solutions in 1993
- 30+ years in business intelligence, dimensional modeling
- Oracle ACE Director    
ACE Director
- Developer for IRI (expert in Oracle OLAP and related)
- BIWA Board Member since 2008
- BA Computer Science Brown University

## Tim Vlami – Vice President & Analytics Strategist

- 30+ years in business modeling and valuation, forecasting, and scenario analyses
- Oracle ACE    
ACE
- Instructor for Oracle University's Data Mining Techniques and Oracle R Enterprise Essentials Courses
- Professional Certified Marketer (PCM) from AMA
- Adjunct Professor of Business Benedictine College
- MBA Kellogg School of Management (Northwestern University)
- BA Economics Yale University



# Vlami Involvement in Presentations

Presenter	Time	Location	Title
Dan Vlami & Mike Caskey	Mon 12:00 PM	Banyan C	Analytic Views Simplify Complex Business Intelligence Queries
Dan Vlami & Mike Caskey	Mon 2:00 PM	Banyan C	Upgrading to Oracle Business Intelligence 12c
Jeff McBride & Mike Caskey	Tues 9:15 AM	Breakers I	Case Study of Improving BI Apps and OBIEE Performance
Dan Vlami & Tim Vlami	Tues 3:30 PM	Banyan C	Oracle Big Data Science
Dan Vlami & Tim Vlami	Wed 9:15 AM	Banyan D	Data Analysis with Various Oracle Business Intelligence and Analytics Tools
Tim Vlami	Thurs 12:15 PM	Jasmine F	BI Movie Magic: Maps, Graphs, and BI Dashboards at AMC Theatres



# Presentation Agenda

- Oracle's Big Data Strategy (our interpretation)
- Oracle Portfolio of Big Data Science Products w/Demos
  - Big Data Discovery
  - Big Data SQL
  - Oracle R Advanced Analytics for Hadoop (ORAAH)
  - Big Data Spatial and Graph
- Predictive Analytics in Hadoop
- Predictive Analytics in Oracle Database (Oracle Advanced Analytics)
  - Oracle Data Mining
  - Oracle R Enterprise
- Strategies for Enterprise Scale Data Science

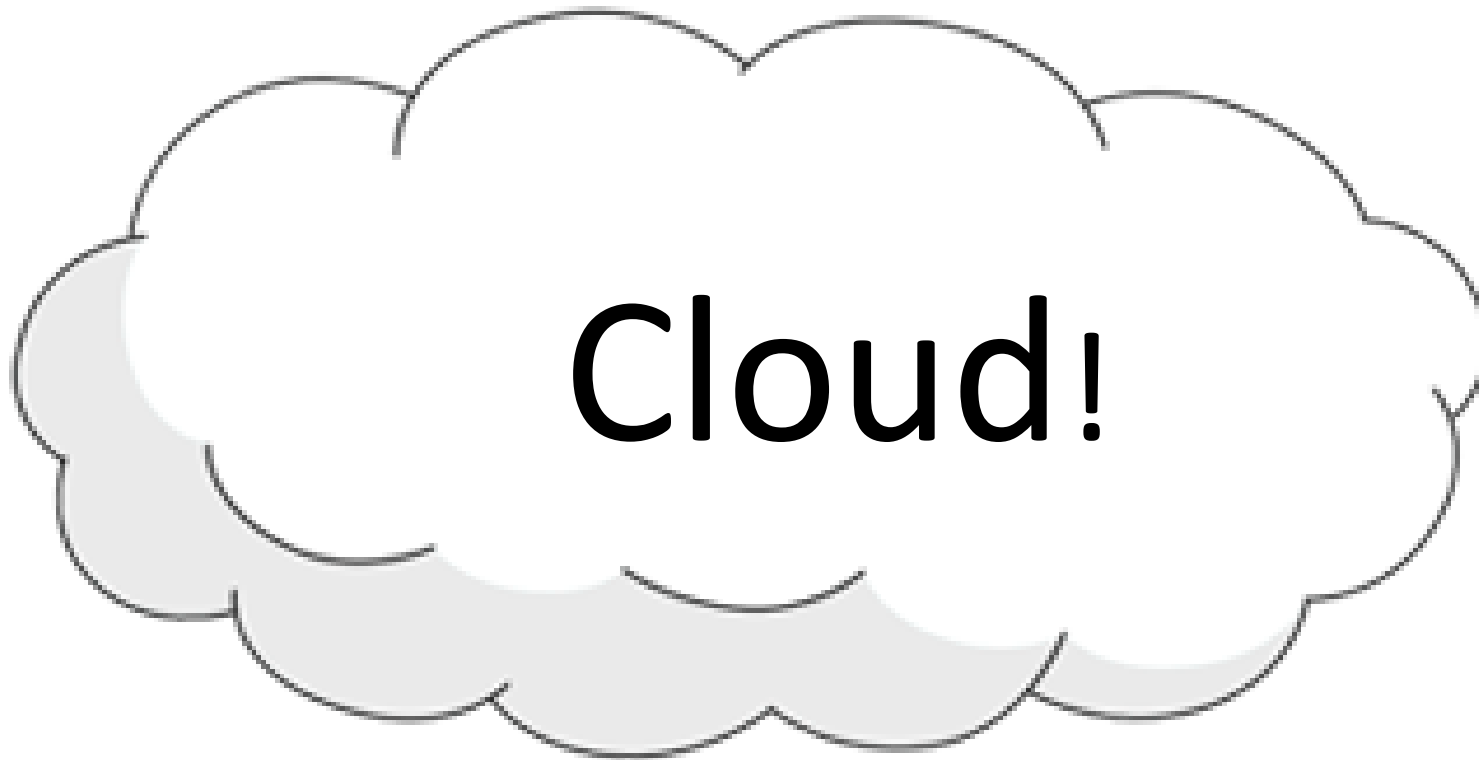


# Oracle's Big Data Strategy

- **Cloud**
  - Big Data Appliance
  - Heterogeneous Data Environments



# Oracle's Big Data Science Strategy





# Many Languages





# Oracle has competing “Service Branches”



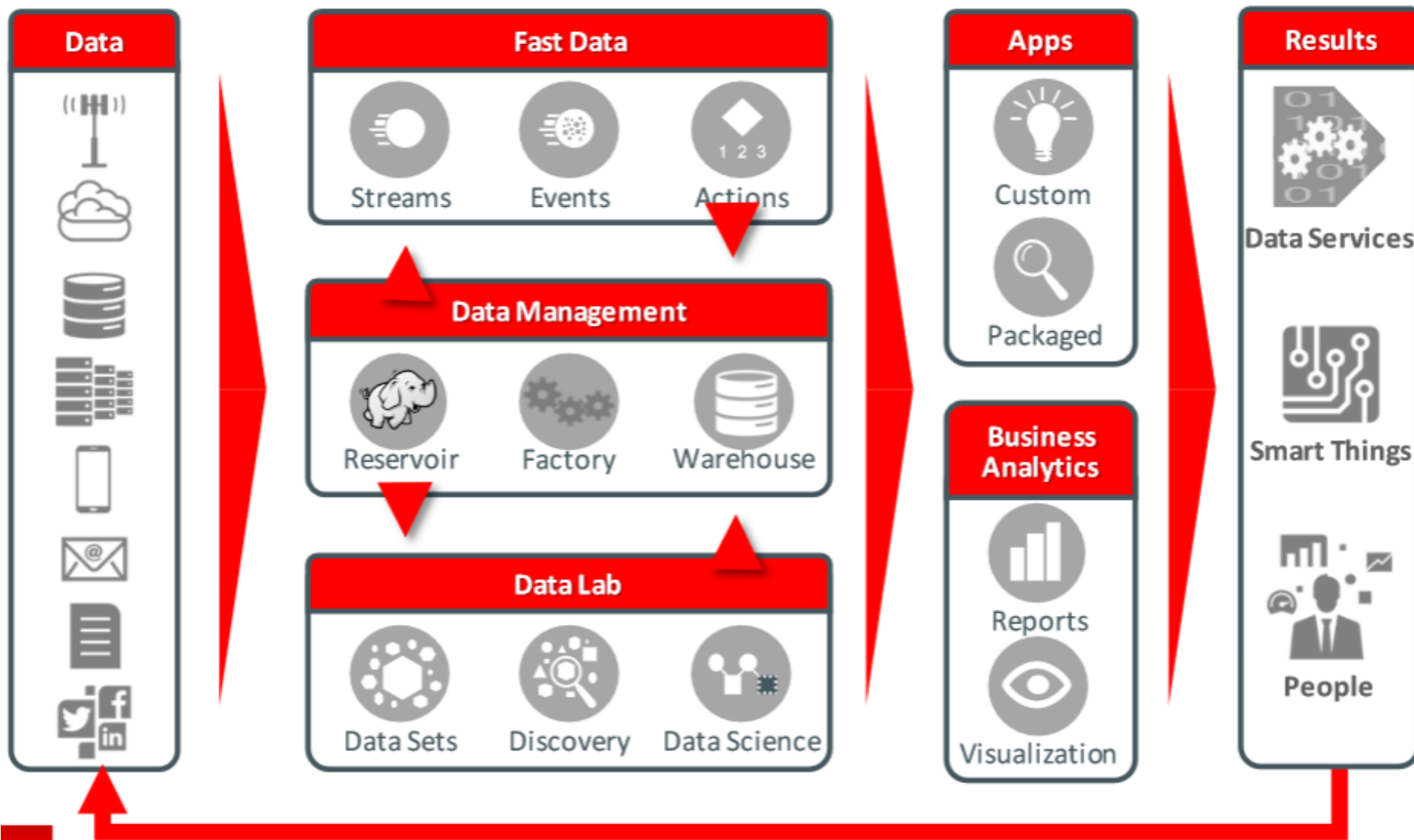


# Oracle's Big Data Strategy

	Per User	Per Processor
Big Data Connectors		\$2,000
Big Data Spatial and Graph		\$2,000
GoldenGate for Big Data	\$400	\$20,000
ODI Advanced Big Data Option	\$150	\$5,000
Big Data Discovery	\$20,000	\$50,000
NoSQL Database EE	\$200	\$10,000
Big Data SQL (BDA only)		\$4,000



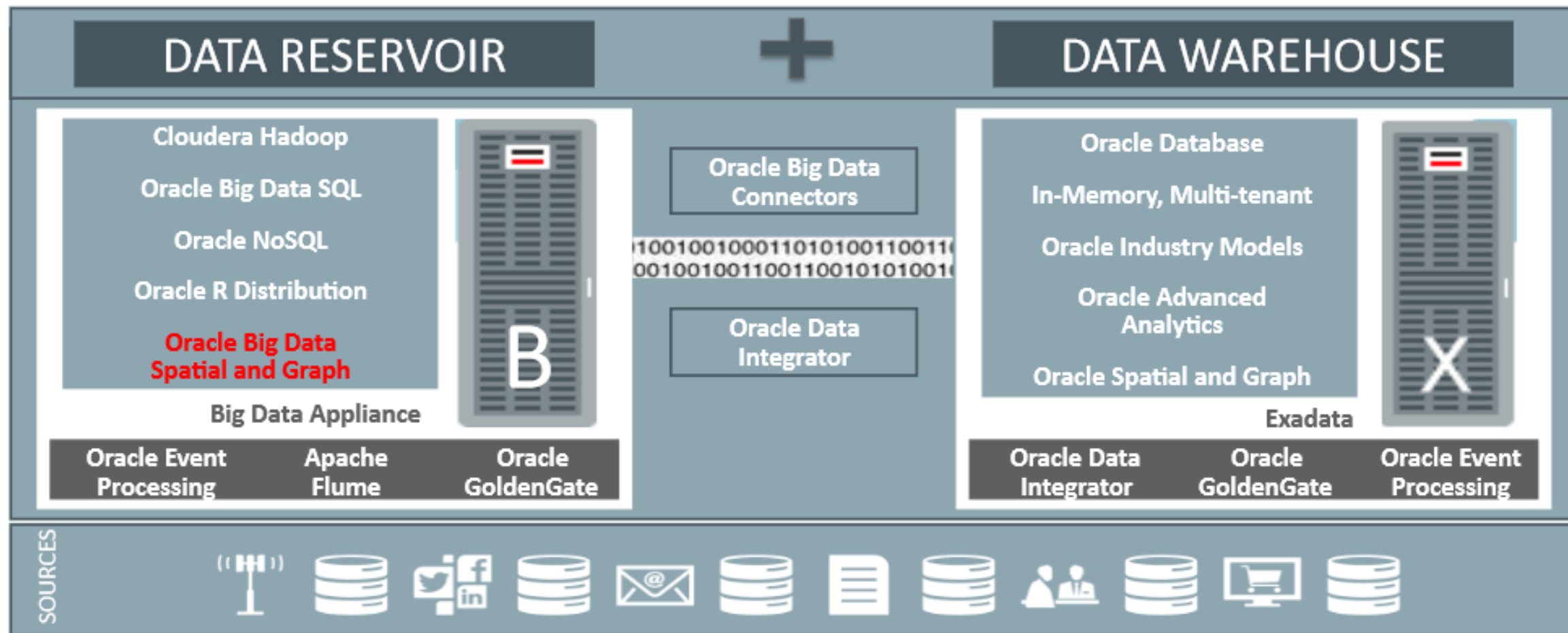
# Big Data Information Flows





# Oracle Big Data Management System

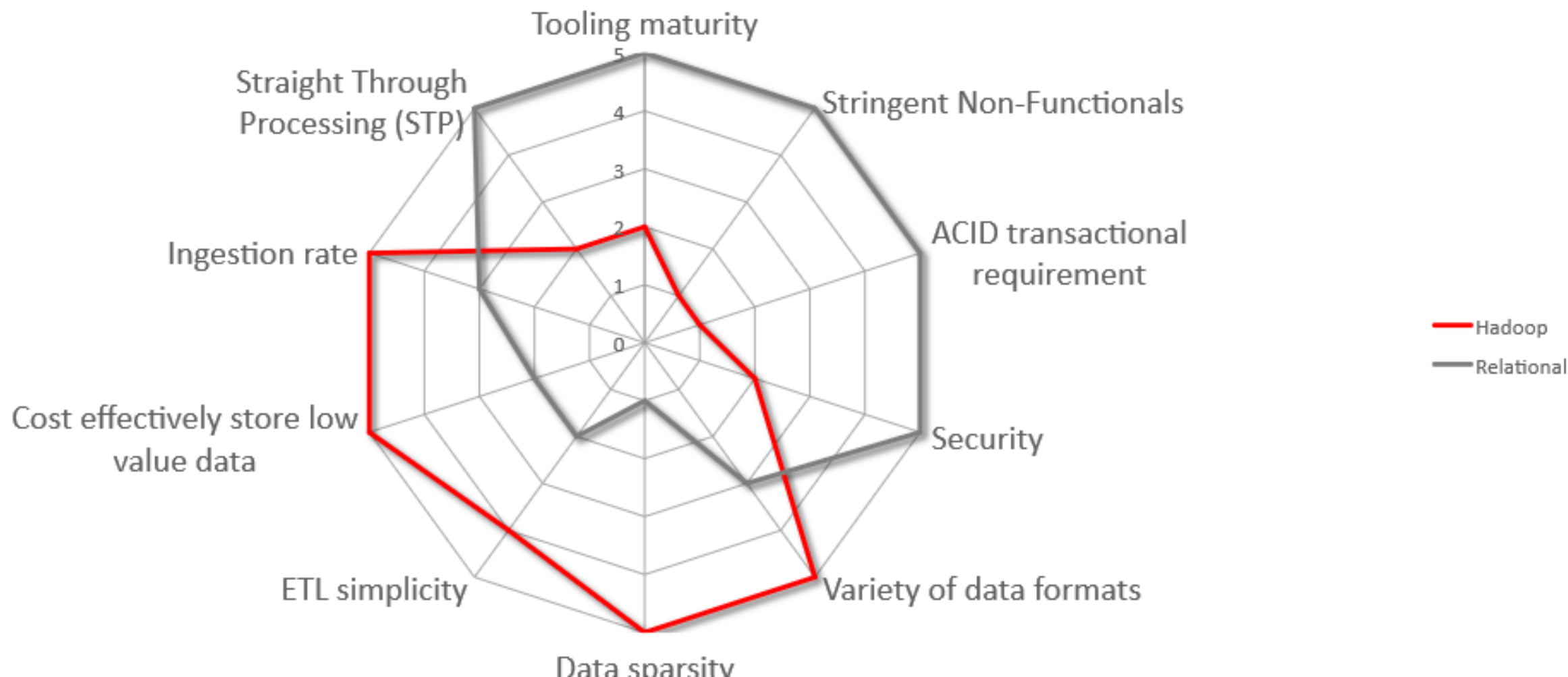
## The Big Picture – Oracle Big Data Management System





# Conventional database or Big Data technologies

## Typical technical decision criteria





# Big Data Science is Young





# Oracle Big Data Discovery

- Visual “front-end” for Hadoop
- Catalog data sets
- Visualize attributes by data type and sort by relevance
- Discover patterns, outliers, and correlations
- Enrich and transform data (munging)
- Explore with interactive visualizations
- Build galleries and tell Big Data Stories



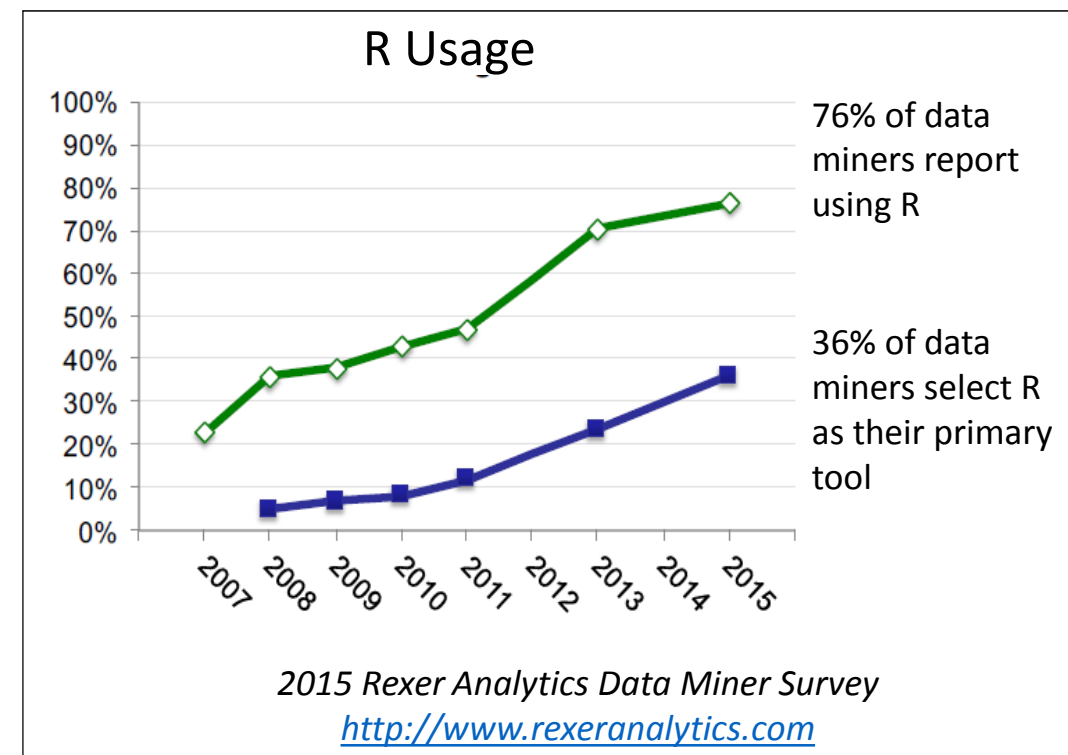
# Big Data Discovery Demo



# What is R?



- An **Open Source** scripting language and environment for statistical computing and graphics <http://www.R-project.org/>
- Popular alternative to **SAS, SPSS** & other proprietary statistical environments
- 2 million+ users worldwide and growing
- Thousands of R packages available
- Taught extensively in higher education



- Oracle R Distribution
- ROracle
- Oracle R Advanced Analytics for Hadoop (ORAAH)
- Oracle R Connector for Hadoop (ORCH)
- Oracle R Enterprise (part of Oracle Advanced Analytics)



# Oracle R Advanced Analytics for Hadoop

- Part of the Big Data Connectors package
  - ORAAH and Oracle Loader for Hadoop
  - Oracle SQL Connector for Hadoop
  - Oracle Xquery for Hadoop
  - Oracle Data Integrator Enterprise Edition (restricted)
- Execution of R scripts in Hadoop
- R interface to Hive tables through transparency layer
- R interface for Oracle database tables through transparency layer
- Set of pre-packaged algorithms
- ORCH (Oracle R Connector for Hadoop)



# ORAAH Algorithms

- Linear regression
- Generalized linear models (GLM)
- Neural Net
- Low rank matrix factorization
- Non-negative matrix factorization
- Principle components analysis (PCA)



# ORAAH MR Hadoop & Spark Functions

Current release 2.5.1



Function	Description
orch.cor	Generates a correlation matrix with a Pearson's correlation coefficients.
orch.cov	Generates a covariance matrix.
orch.getXlevels	Creates a list of factor levels that can be used in the xlev argument of a model.matrix call. It is equivalent to the .getXlevels function in the stats package.
orch.glm	Fits and uses generalized linear models on data stored in HDFS. Can fit the algorithm using the new Spark-based computation for a much faster computation and scoring as well.
orch.kmeans	Perform k-means clustering on a data matrix that is stored as a file in HDFS.
orch.lm	Fits a linear model using tall-and-skinny QR (TSQR) factorization and parallel distribution. The function computes the same statistical parameters as the Oracle R Enterprise ore.lm function.
orch.lmf	Fits a low rank matrix factorization model using either the jellyfish algorithm or the Mahout alternating least squares with weighted regularization (ALS-WR) algorithm.
spark.connect	Connects to an Apache Spark server through YARN or Standalone mode and creates a Context for use with the Spark-based algorithms in ORAAH

ORACLE

Copyright © 2016 Oracle and/or its affiliates. All rights reserved. |

70



# ORAAH MR Hadoop & Spark Functions

Current release 2.5.1



Function	Description
orch.neural	Provides a highly scalable Multi-Layer Perceptron Neural Network to model complex, nonlinear relationships between inputs and outputs, or to find patterns in the data. Automatically detects connection to a Spark Context to use Spark for faster computation.
orch.nmf	Provides the main entry point to create a nonnegative matrix factorization model using the jellyfish algorithm. This function can work on much larger data sets than the R NMF package, because the input does not need to fit into memory.
orch.nmf.NMFalgo	Plugs in to the R NMF package framework as a custom algorithm. This function is used for benchmark testing.
orch.princomp	Analyzes the performance of principal component.
orch.recommend	Computes the top $n$ items to be recommended for each user that has predicted ratings based on the input orch.mahout.lmf.asl model.
orch.sample	Provides the reservoir sampling.
orch.scale	Performs scaling.

ORACLE®

Copyright © 2016 Oracle and/or its affiliates. All rights reserved. |

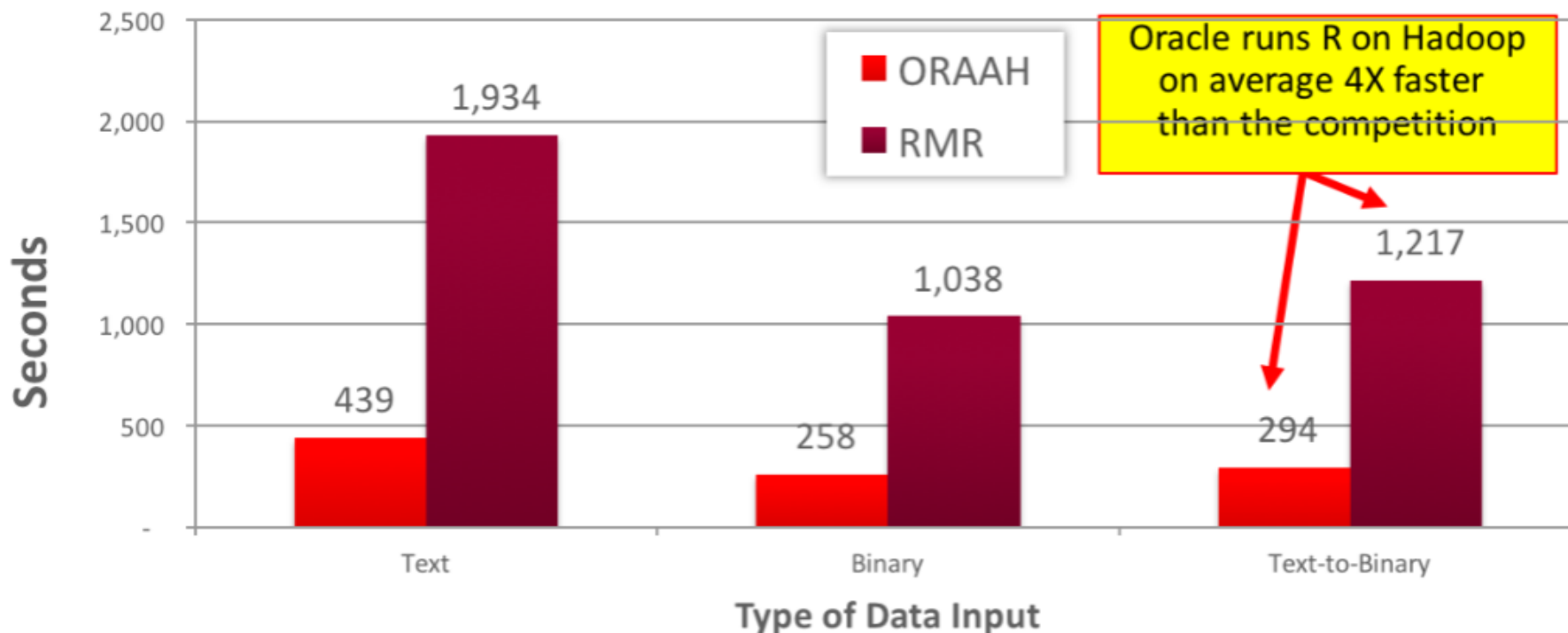


# ORAAH is Fast

## Oracle R Advanced Analytics for Hadoop – vs. Rhadoop (RMR)

Best platform available to run Hadoop-R jobs vs. Revolution Analytics' RHadoop

Performance on a 6-node BDA X3-2, 16 cores and 47 GB of Total RAM assigned  
Covariance computation on 100 GB HDFS/200 columns input dataset



[https://blogs.oracle.com/R/entry/oraah\\_enabling\\_high\\_performance\\_r](https://blogs.oracle.com/R/entry/oraah_enabling_high_performance_r)



# Oracle Advanced Analytics

- Licensed option to Oracle Database Enterprise Edition
- Two primary components
  - Oracle Data Mining
  - Oracle R Enterprise
- Includes an extensive set of APIs, algorithms, and capabilities

# Oracle's Advanced Analytics

In-Database Data Mining Algorithms\*—SQL &  & GUI Access



## Classification



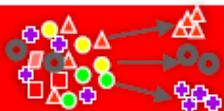
- Decision Tree
- Logistic Regression (GLM)
- Naïve Bayes
- Support Vector Machine (SVM)
- Random Forest

## Regression



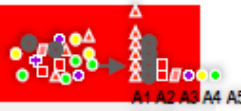
- Multiple Regression (GLM)
- Support Vector Machine (SVM)
- Linear Model
- Generalized Linear Model
- Multi-Layer Neural Networks
- Stepwise Linear Regression

## Clustering



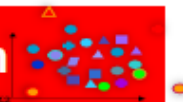
- Hierarchical k-Means
- Orthogonal Partitioning Clustering
- Expectation-Maximization

## Attribute Importance



- Minimum Description Length
- Unsupervised pair-wise KL div.

## Anomaly Detection



- 1 Class Support Vector Machine

## Time Series

- Single & Double Exp. Smoothing

## Predictive Queries

- Clustering
- Regression
- Anomaly Detection
- Feature Extraction

## Feature Extraction & Creation

- Nonnegative Matrix Factorization
- Principal Component Analysis
- Singular Value Decomposition

## Market Basket Analysis



- Apriori – Association Rules

## Open Source R Algorithms

- Ability to run any R package via Embedded R mode



\* supports partitioned models, text mining



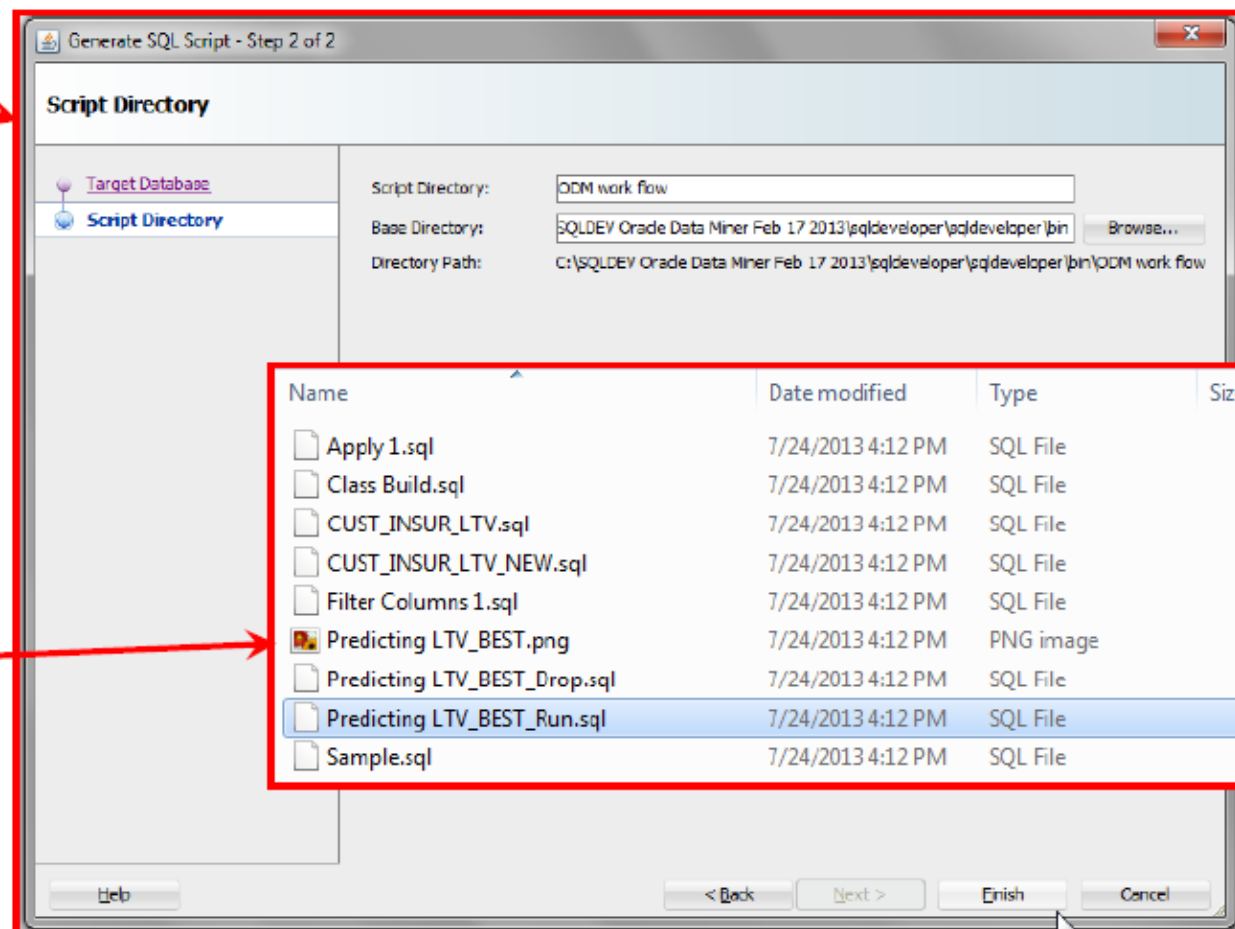
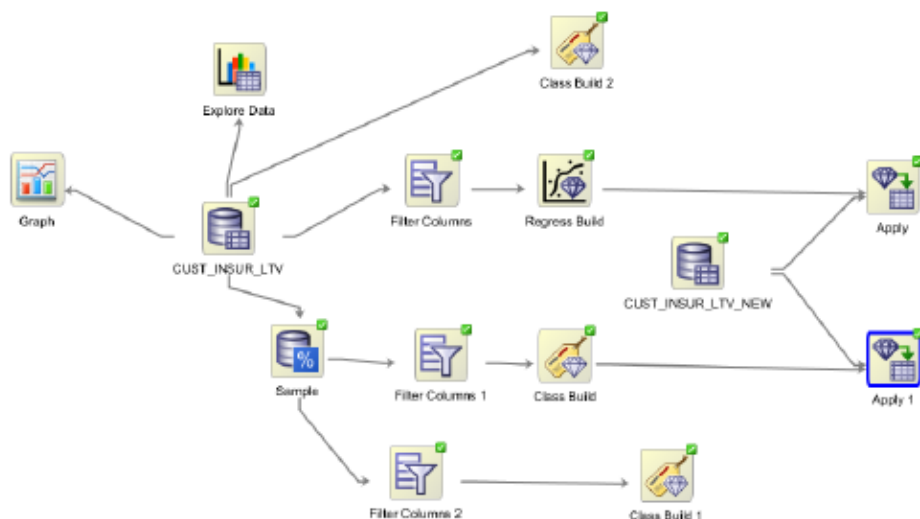
# Data Miner GUI for Analytic Workflows

## SQL Developer/Oracle Data Miner 4.0 New Features



### ■ SQL Script Generation

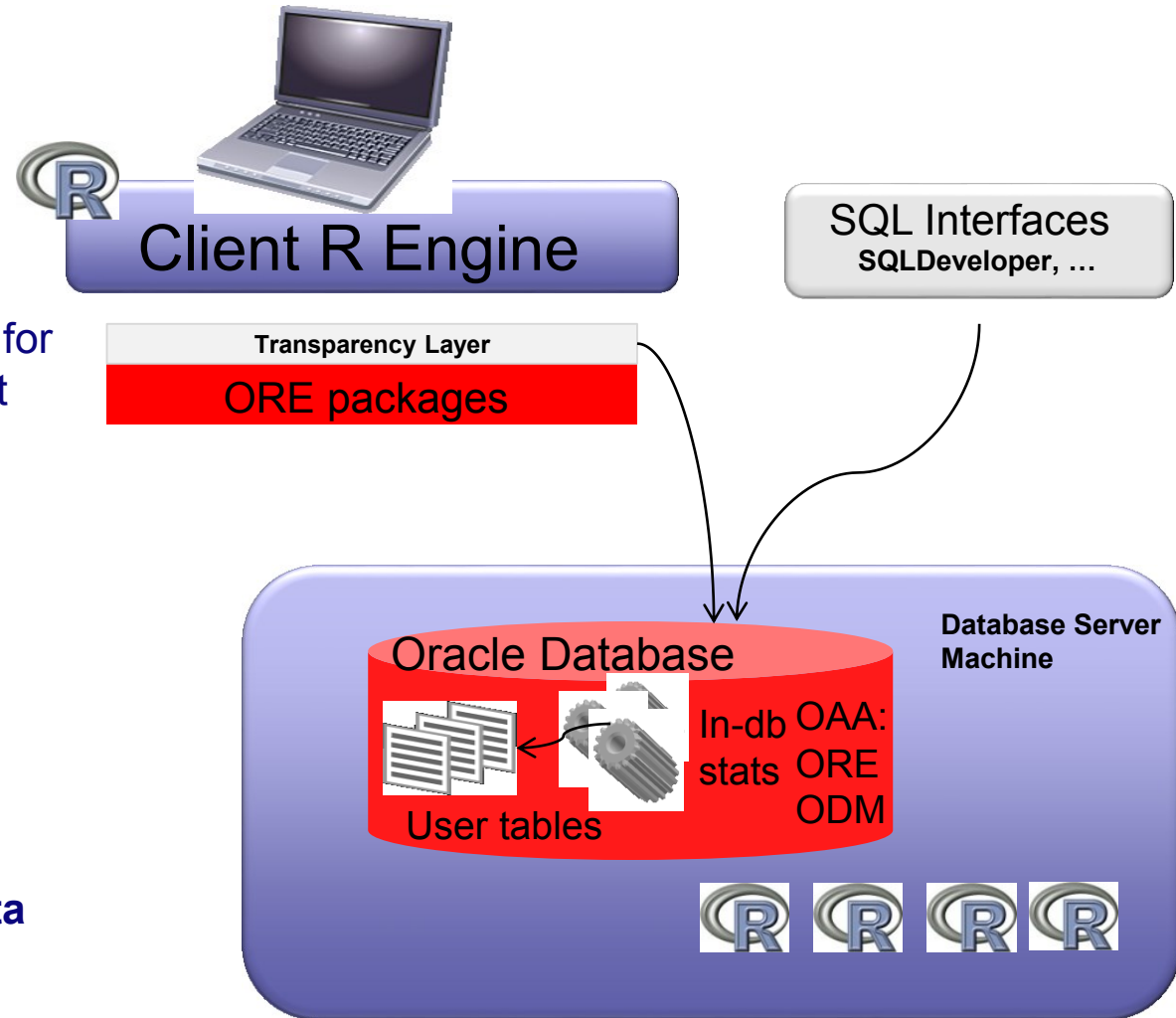
- Deploy entire methodology as a SQL script
- Immediate deployment of data analyst's methodologies





# Oracle R Enterprise

- A comprehensive, database-centric environment for end-to-end analytical processes in R, with immediate deployment to production environments
- Operationalize entire R scripts in production applications – eliminate porting R code
- Seamlessly leverage Oracle Database as an HPC environment for R scripts, providing data parallelism and resource management
- Avoid reinventing code to integrate R results into existing applications
- Transparently analyze and manipulate data in Oracle Database through R using versatile and customizable R functions
- Eliminate memory constraint of client R engine
- Score R models in Oracle Database
- Execute R scripts through Oracle Database server machine for scalability and performance
- **Get maximum value from your Oracle Database and Exadata**
- Enable integration and management through SQL
- Integrate R into the IT software stack, e.g. OBIEE





# Sensor Data Analysis

- 200K households, each with a utility “smart meter”
- 1 reading/meter/hour
- 200K x 8760 hours/year → 1.752B readings per year
- 3 years worth of data → 5.256B readings
- Each customer has 26280 readings
- Build one model per customer to understand/predict customer monthly usage
- If each model takes 10 seconds to build, 556 hours (23+ days)  
...with 128 DOP → 4.4 hours





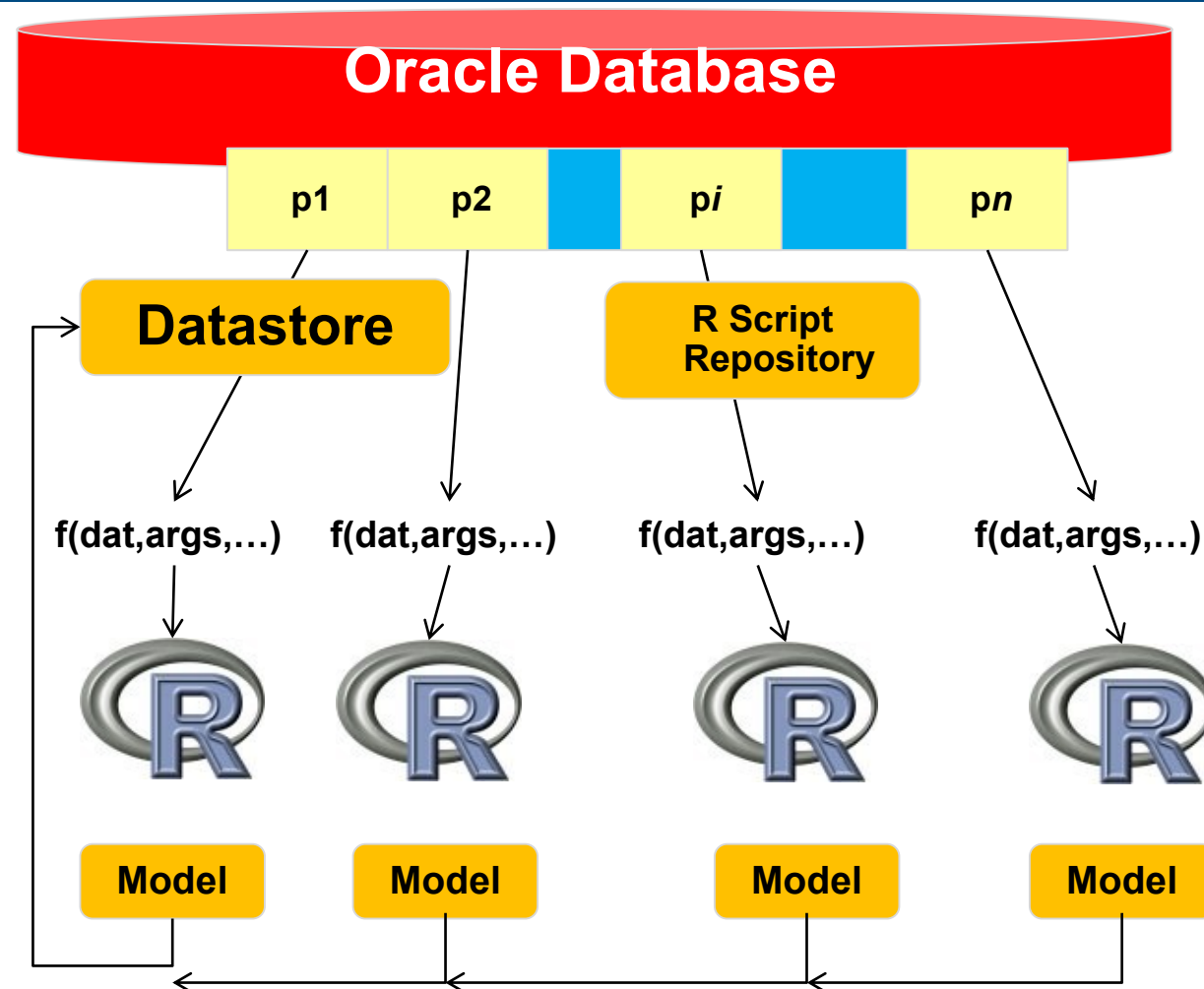
# Smart Meter scenario



$f(\text{dat}, \text{args}, \dots)$  {

R Script  
build  
model

}

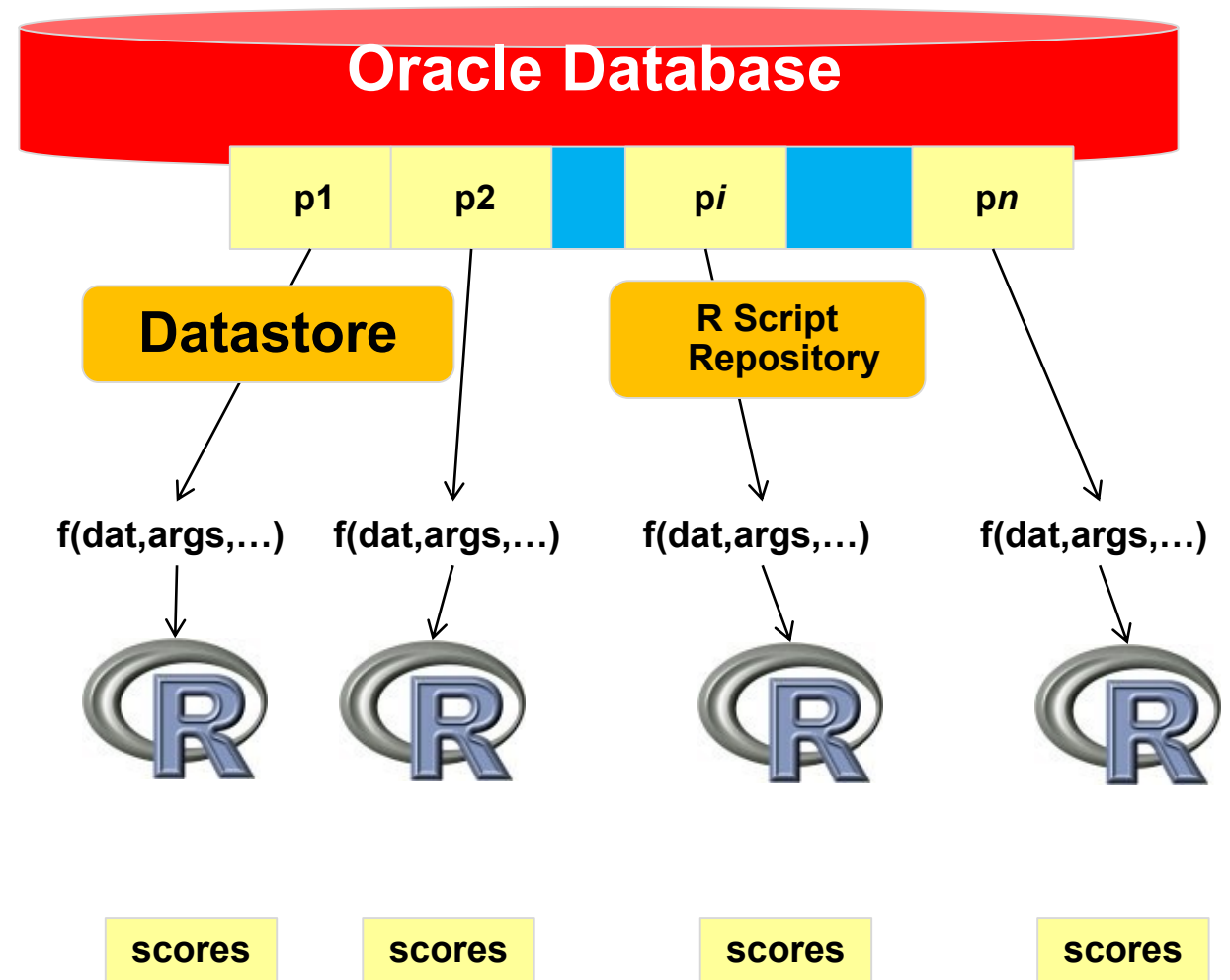




# Smart Meter scenario



```
f(dat,args,...) {  
  R Script  
  score  
  data  
}
```





# Oracle Big Data Connectors

- Oracle SQL Connector for HDFS
  - (previously called Oracle Direct Connector for HDFS)
  - Enables an Oracle external table to access data stored in HDFS or HIVE
  - Data can remain in source or be loaded into Oracle Database
- Oracle Loader for Hadoop
  - High-performance loader for HDFS data into Oracle Database
  - Transforms data in DB format and can sort by Pkey or user defined columns
- Oracle Xquery for Hadoop
  - Runs transformations expressed in XQuery by translating into MapReduce
- Oracle Data Integrator (limited license)
  - ETL tool w/ GUI for data movement to and from ODM, 3<sup>rd</sup> party, and Hadoop
- ORAAH (Oracle Advanced Analytics for Hadoop)



# Big Data Connectors Certification Matrix

[Overview](#)[Features & Benefits](#)[Resources](#)[Certifications](#)

## Oracle Big Data Connectors Certification Matrix

	Oracle SQL Connector for HDFS	Oracle Loader for Hadoop	Oracle Data Integrator Enterprise Edition *	Oracle XQuery for Hadoop	Oracle R Advanced Analytics for Hadoop	Certified by	Support
CDH 4.x (Cloudera)	●	●	●	●	●	Oracle	Oracle
CDH 5.x (Cloudera)	●	●	●	●	●	Oracle	Oracle
Apache Hadoop 2.x	●	●	●	●	●	Oracle	Oracle
HDP 1.3 (Hortonworks)	●	●	●	●	●	Hortonworks	Oracle supports connectors, Customer contacts Hortonworks directly for Hadoop specific issues
HDP 2.1 (Hortonworks)	●	●	●	●	●	Hortonworks	Oracle supports connectors, Customer contacts Hortonworks directly for Hadoop specific issues



# Big Data Spatial and Graph

- Property Graph Capability
- 35 high-performance analytic functions
  - Network theory analytics: centrality, connectedness, betweenness, etc.
- Text search integration through Lucene/SOLR
- Java APIs include
  - TinkerPop/Gremlin
  - Blueprints
  - Hadoop
  - NoSQL
  - HBase



# Oracle University OAA Courses

- Oracle Database 11g: Data Mining Methods

May 9-12, 2016 Chicago, IL

- Predictive Analytics Using Oracle Data Mining (12c)
- Oracle R Enterprise Essentials



# Thank You!

## Oracle Big Data Science

Session 4762

Tim Vlami

[tvlamis@vlamis.com](mailto:tvlamis@vlamis.com)

[www.vlamis.com](http://www.vlamis.com)