



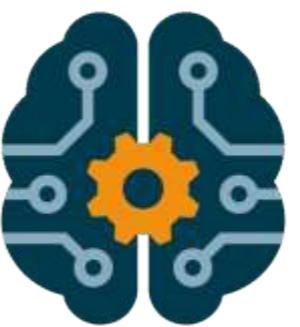
セッション1

オラクルのRテクノロジーの概要

(Oracle Machine Learningを使用)

Senior Director、Mark Hornick
Oracle Machine Learning製品管理

2020年11月



免責条項

下記事項は、弊社の一般的な製品の方向性に関する概要を説明するものです。また、情報提供を唯一の目的とするものであり、いかなる契約にも組み込むことはできません。マテリアルやコード、機能の提供をコミットメント（確約）するものではなく、購買を決定する際の判断材料になさらないでください。オラクル製品に関して記載されている機能の開発、リリースおよび時期については、弊社の裁量により決定されます。

アジェンダ

- 1 Rとは
- 2 Oracle Machine Learningの概要
- 3 Oracle RDistribution
- 4 ROracleパッケージ
- 5 Oracle Machine Learning for SparkOracle
- 6 Machine Learning for R
- 7 まとめ

Rとは

Rは、統計コンピューティングとグラフィック向けのオープンソースのスクリプト言語および環境 (<http://www.R-project.org/>)

1994年にSAS、SPSS、その他の専有の統計環境に代わる選択肢として開発

データ操作、計算、グラフィカル表示向けのソフトウェア機能の統合スイート

Rユーザーは世界中で数百万人

- 大学で広く教えられている
- 多くの企業アナリストとデータ・サイエンティストがRを認識および使用

以下をはじめとする、生産性を向上するための数千のオープンソース・パッケージが存在

- Bioinformatics with R
- Spatial Statistics with R
- Financial Market Analysis with R
- Linear and Non Linear Modeling

Topics

Bayesian	ベイズ推定
ChemPhys	計量化学と計算物理学
ClinicalTrials	臨床試験の設計、監視、分析
Cluster	クラスタ分析と有限混合モデル
Databases	Rを使用したデータベース
DifferentialEquations	微分方程式
Distributions	確率分布
Econometrics	計量経済学
Environmetrics	生態学的データと環境データの分析
ExperimentalDesign	実験計画法（DoE）と実験データの分析
ExtremeValue	極値解析
Finance	経験的ファイナンス
FunctionalData	関数データ解析
Genetics	遺伝統計学
Graphics	グラフィック表示、動的グラフィック、グラフィック・デバイス、可視化
HighPerformanceComputing	Rを使用した高パフォーマンスの並列コンピューティング
Hydrology	水文学的データおよびモデリング
MachineLearning	機械学習と統計学習
MedicalImaging	医療画像分析
MetaAnalysis	メタ分析
MissingData	欠損データ
ModelDeployment	Rを使用したモデル開発
Multivariate	多変量解析
NaturalLanguageProcessing	自然言語処理
NumericalMathematics	数値数学
OfficialStatistics	正式な統計および調査手法
Optimization	最適化と数理プログラミング
Pharmacokinetics	薬物動態学データの分析
Phylogenetics	系統発生学、特に比較研究法
Psychometrics	精神測定モデルおよび手法
ReproducibleResearch	再現可能な研究
Robust	堅牢な統計手法
SocialSciences	社会科学向けの統計
Spatial	空間データの分析
SpatioTemporal	空間および時間データの扱いと分析
Survival	生存分析
TeachingStatistics	統計の授業
TimeSeries	時系列分析
Tracking	追跡データの処理と分析
WebTechnologies	Webテクノロジーおよびサービス
gR	Rでのグラフィカル・モデル



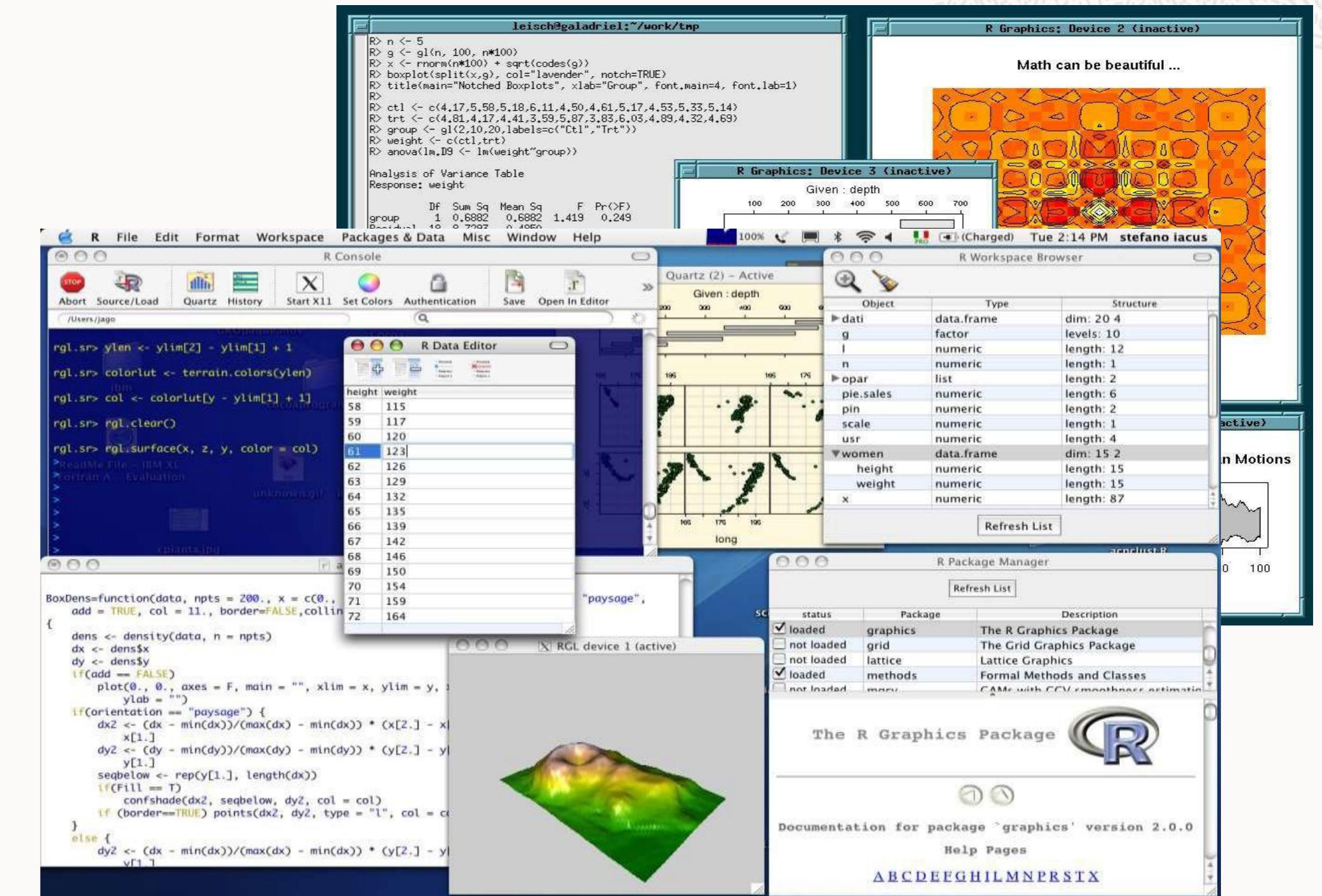
データ・サイエンティスト、統計学者、データ・アナリストがRを使用する理由

Rは、Base SASやSPSS Statisticsと類似した統計言語

R環境の特徴

- パワフル
- 拡張可能
- グラフィカル
- 多彩な統計
- 多くの‘つまみ’とスマートなデフォルトを備えたOOTB機能
- 容易なインストールと使いやすさ
- 無償

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



分析における課題

データを得る、つまり‘適切な’データを得るのに時間がかかりすぎる
すべてのデータを分析または処理できないため、サンプリングが必要である分析/
予測モデルを作成し、結果を本番環境に適用するのは場当たり的で複雑である
Rまたはその他のモデルをSQL、C、Javaに再コード化するのは時間がかかり、エラーを引き起こ
しやすいデータのセキュリティ、バックアップ、リカバリについて懸念している
ビジネス目標を達成するには、何万ものモデルを迅速に構築する必要がある

次のブログ・シリーズをお読みください：

https://blogs.oracle.com/R/entry/addressing_analytic_pain_points

Oracle Machine Learning

Oracle Machine Learningの差別化要因

データベースとHadoopで直接データを操作

IT/DBAに抽出を依頼する必要性をなくし、データベースとHadoopのデータに即座にアクセスデータをそのデータがある場所で処理することで、データの移動を最小化または排除

スケーラビリティとパフォーマンス

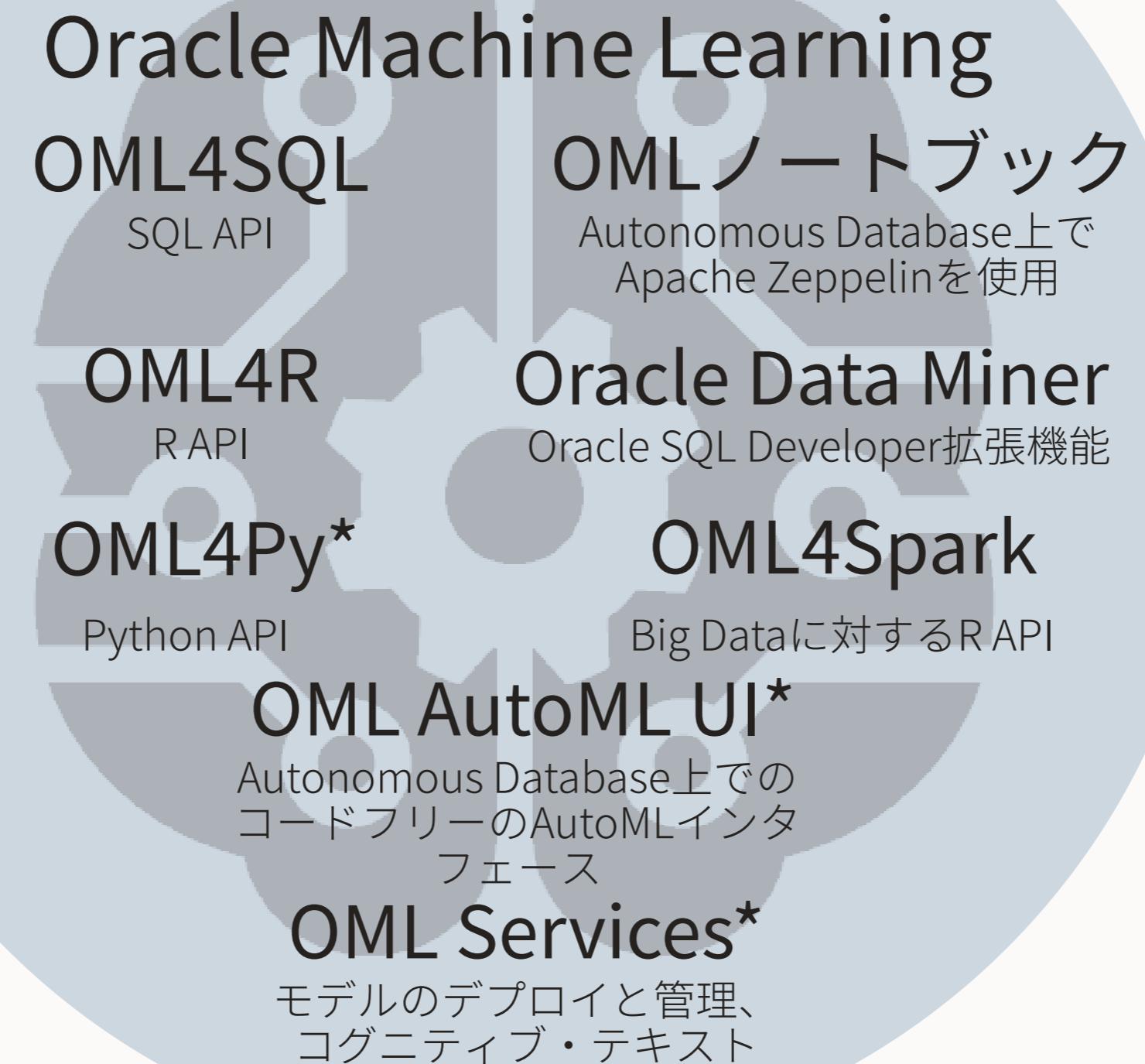
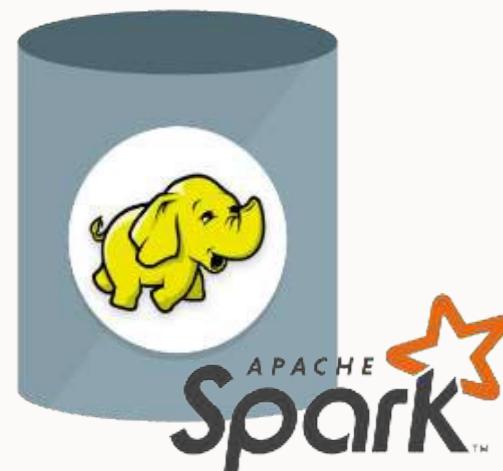
Oracle Databaseでビッグ・データにスケーリングする並列分散アルゴリズムを使用Exadataクラスのマシンを活用して数十億のデータ行でモデルを構築

容易なデプロイメント

Oracle Databaseを使用して、R、Python、およびSQLのスクリプトを本番環境に即座にデプロイ（再コード化は不要）カスタム構築が不要で余計な複雑性のない本番品質のインフラストラクチャを使用

プロセスのサポート

既存のプロセスを使用してデータのセキュリティ、バックアップ、リカバリを確保および保守
Oracle Databaseでの分析オブジェクト（モデル、スクリプト、ワークフロー、データ）の保管、アクセス、管理、追跡を実現



* 近日追加予定



Oracle DatabaseへのOracle Machine Learningインターフェース

ツール Oracle Machine Learningコンポーネント データ管理プラットフォーム



Apache Zeppelin



Pythonクライアント、
Jupyter Notebook



Oracle SQL Developer
Oracle SQL*Plus



Rクライアント、
RStudio



SQL Developer

OMLノートブック

OML4Py*

OML4SQL

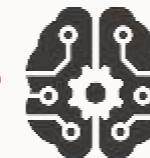
OML4R

Oracle Data Miner

OML4SQL

OML4Py*

OML4R*



自律型データベース



Oracle Database



データベース
Cloud Service

Oracle Machine Learningのアルゴリズムと分析

分類

- ナイーブ・ベイズ
- ロジスティック回帰 (GLM)
- ディシジョン・ツリー
- ランダム・フォレスト
- ニューラル・ネットワーク
- サポート・ベクター・マシン (SVM)
- 明示的セマンティック分析
- XGBoost*

異常検出

- One-Class SVM
- MSET-SPRT*

クラスタリング

- 階層型k平均法
- 階層型O-Cluster
- 期待値最大化 (EM)

時系列

- 予測 - 指数平滑法
- 一般的なモデルを含む
例: トレンド、季節性、不規則性、欠損データを扱うHolt-Winters

回帰

- 線形モデル
- 一般化線形モデル (GLM)
- サポート・ベクター・マシン (SVM)
- ステップワイズ線形回帰
- ニューラル・ネットワーク
- LASSO
- XGBoost*

属性の重要度

- 最小記述長
- 主成分分析 (PCA)
- 教師なしペアワイズKLダイバージェンス
- 行およびAIのCUR分解

相関ルール

- アプリオリ/マーケット・バスケット

予測問合せ

- 予測、クラスタ、検出、特徴

SQL分析

- SQLウィンドウ
- SQLパターン
- SQL集計

特徴抽出

- 主成分分析 (PCA)
- Non-negative Matrix Factorization
- 特異値分解 (SVD)
- 明示的セマンティック分析 (ESA)

行の重要度

- CUR分解

ランキング

- XGBoost*

テキスト・マイニングのサポート

- アルゴリズムでテキスト列をサポート
- トクナイゼーションとテーマ抽出
- 明示的セマンティック分析 (ESA)

統計関数

- 最小、最大、中央値、標準偏差、t検定、F検定、ピアソンのカイ二乗検定、分散分析、その他

RおよびPythonのパッケージ

- 組込み実行によるサード・パーティのRおよびPythonのパッケージ
- Spark MLlibアルゴリズム統合

Oracle Machine Learning Notebooks

データ・サイエンス・プラットフォームとしてのAutonomous Database

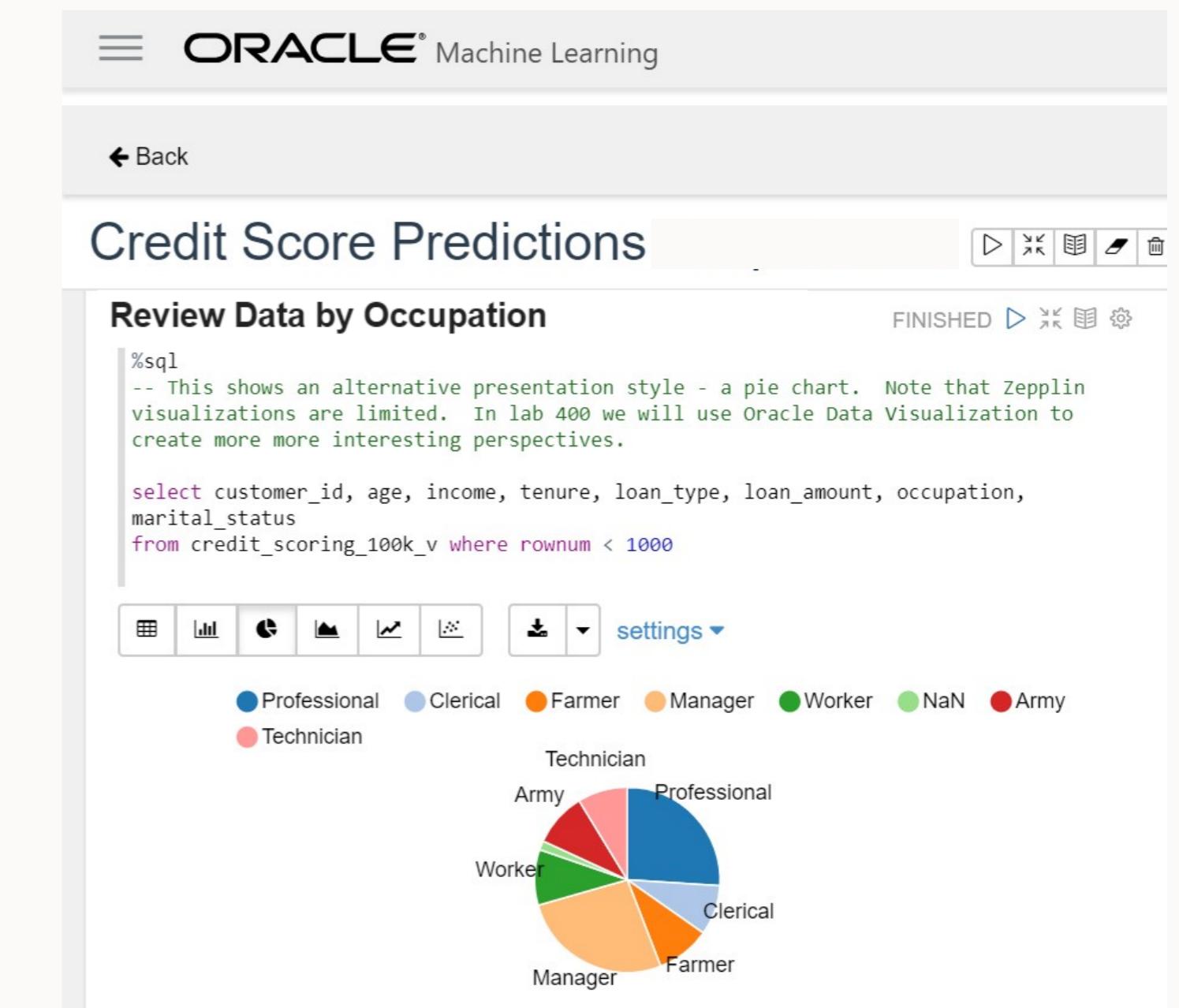


共同作業に適したUI

- Apache Zeppelinが基盤
- SQLとPythonを使用するデータ・サイエンティスト、データ・アナリスト、アプリケーション開発者、DBAをサポート
- ノートブックとテンプレートの容易な共有
- アクセス権、バージョニング、実行スケジューリング

Autonomous Databaseに搭載

- 自動のプロビジョニング、管理、バックアップ
- インデータベース・アルゴリズムおよび分析関数
- モデルの調査、準備、構築、評価、データのスコアリング、ソリューションのデプロイ
- 近日中にRによる補強を予定



Oracle Machine Learning for SQL

Oracle DatabaseおよびOracle Autonomous Database内の
MLにすぐにアクセスできるようSQLユーザーをサポート

インデータベースの並列分散アルゴリズム

- MLエンジンを分離するためのデータ抽出が不要
- 高速でスケーラブル
- バッチおよびリアルタイムのスコアリング
- 分かりやすい予測の詳細

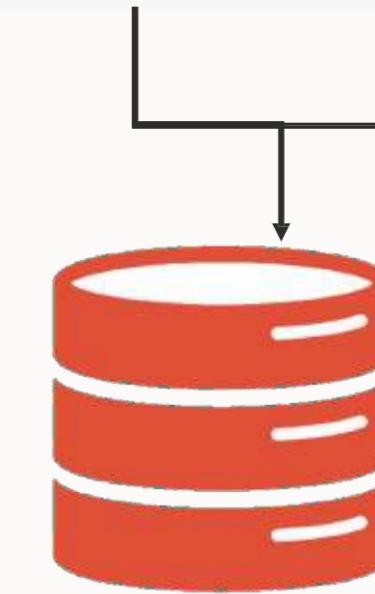
もっとも重要なデータベース・オブジェクトとしてのMLモデル

- アクセス権を使用したアクセス制御
- ユーザー・アクションの監査
- データベース間のモデルのエクスポート/インポートOracle
スタック全体でのMLの活用

SQLインターフェース
SQL*Plus
SQL Developer
...



OMLノートブック



OMLを使用した
Oracle Database

Oracle
Autonomous
Database

Oracle Data Minerのユーザー・インターフェース

分析ワークフローの作成 - データ・サイエンティスト向けの生産性ツール -
専門家ではなくてもデータ・サイエンティストに

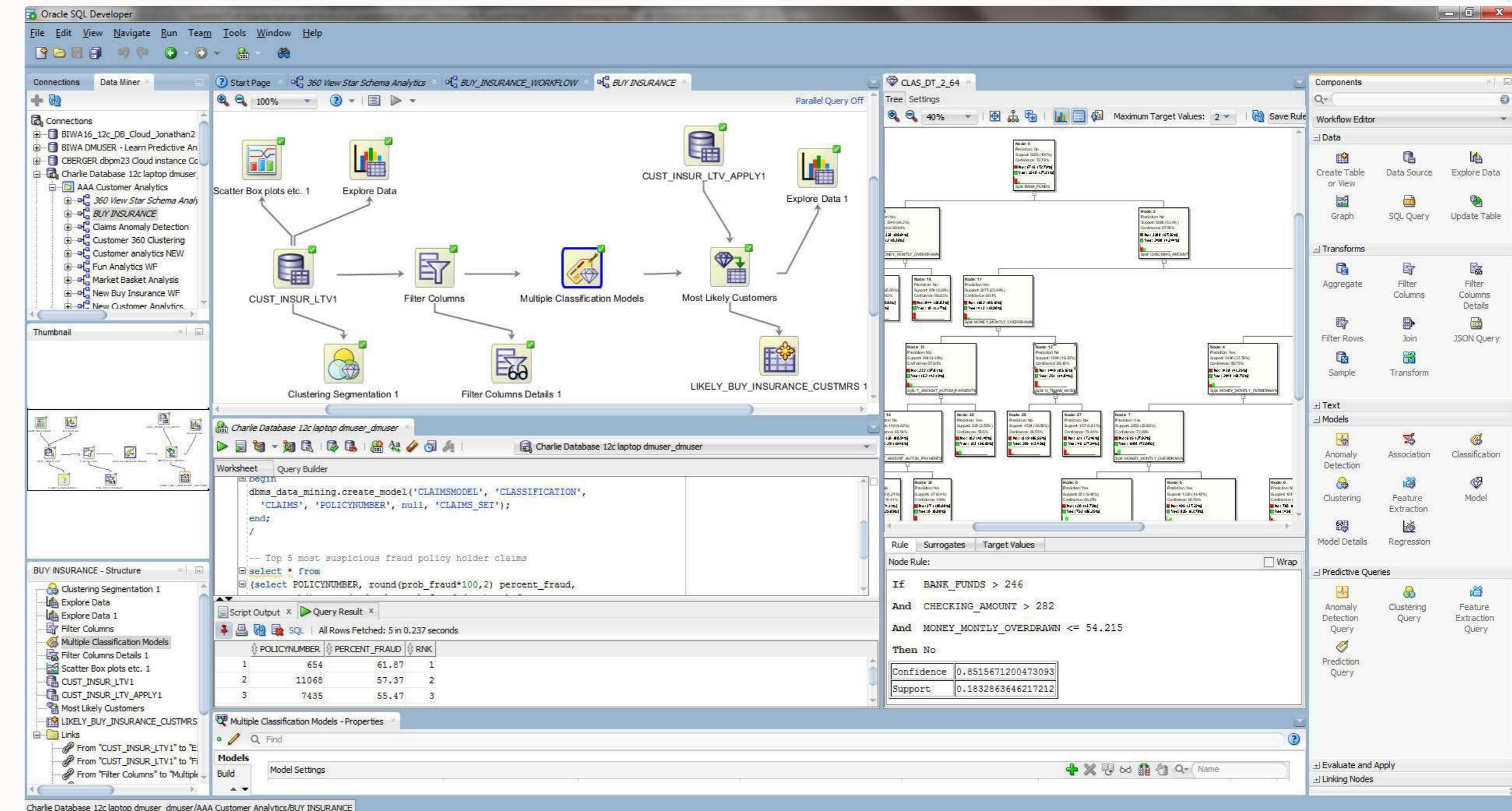
オンプレミスおよびDBCS
のOracle Database用のSQL
Developer拡張機能

標準的なデータ・サイエ
ンスの手順を自動化
使いやすいドラッグアンドド
ロップ・インターフェース

分析ワークフローを迅速に定
義、共有

幅広いアルゴリズムとデータ
変換処理

即時デプロイメント用の
SQLコードを生成



Oracle Machine Learning for R、Python

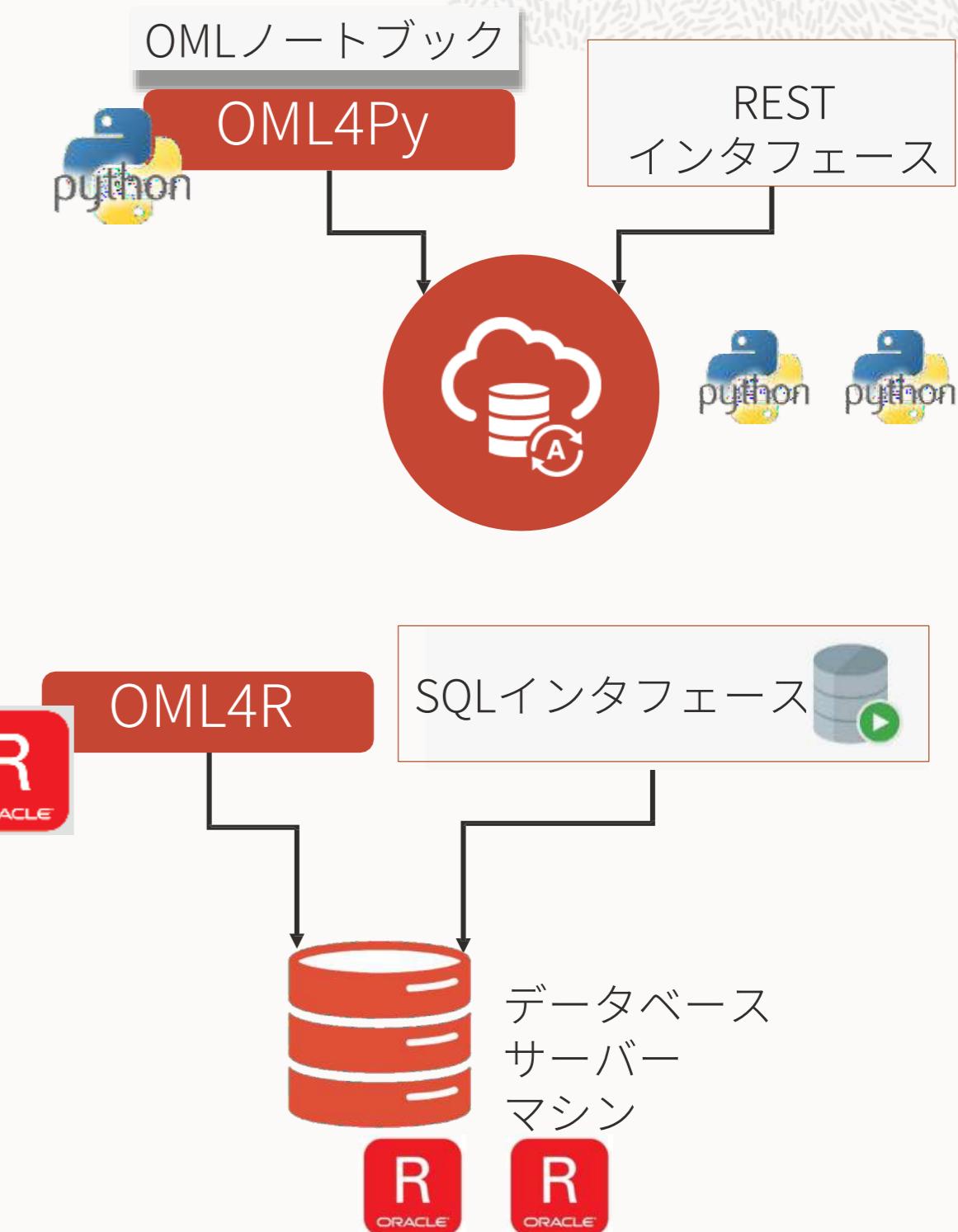
オープン・ソース環境においてデータ・サイエンティストをサポート

HPC環境としてのOracle Database
並列で分散されたインデータベースの
機械学習アルゴリズムを使用

Oracle Database内でスクリプトとオブジェクトを管理

結果をSQLまたはREST経由で
アプリケーションおよびダッシュボードに統合

OML4Pyの自動機械学習



ユースケース例

顧客のトランザクションや保険金請求での不正検出

特定の症状を発症するリスクがある患者の特定適切なオファーを使用した適切な顧客の絞り込み

隠れた顧客セグメントの発見

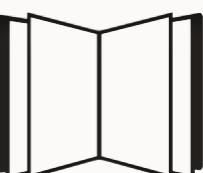
製品またはサービスに対する顧客の需要の予測

もっとも収益の多い販売機会の発見顧客離れの予測と防止他社に流れる可能性のある顧客とその理由の特定

セキュリティと疑わしいアクティビティの検出顧客の会話におけるセンチメントの理解

ソーシャル・ネットワークにおけるインフルエンサーの把握

信用リスクの予測



オラクルのRテクノロジー

R、Oracle Database、Oracle Big Data Appliance/Hadoopをサポート

Oracle R Distribution

ROracle

Rコミュニティが無償で
利用できるソフトウェア

Oracle Machine Learning for R

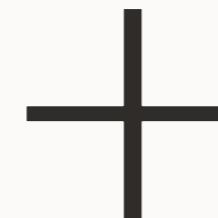
Oracle DatabaseライセンスおよびOracle Database Cloud Serviceと同梱

Oracle Machine Learning for Spark

Oracle Big Data Connectorsソフトウェア・スイートとOracle Big Data Serviceの
コンポーネント

Oracle R Distribution

Oracle R Distribution

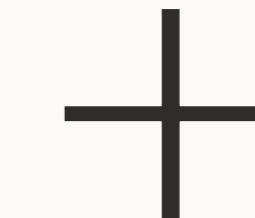


動的なロード機能

Intel Math Kernel Library

AMD Core Math Library

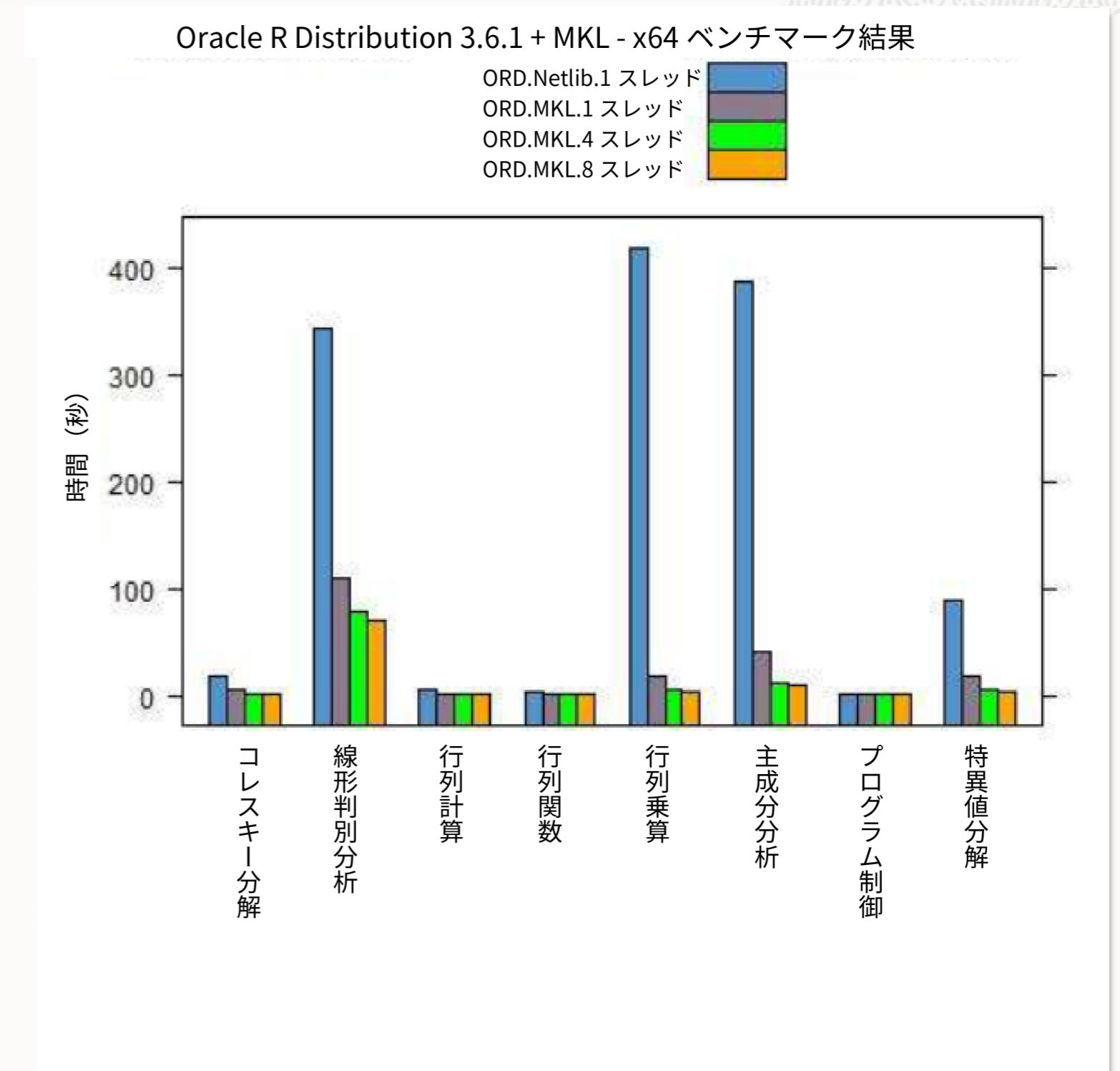
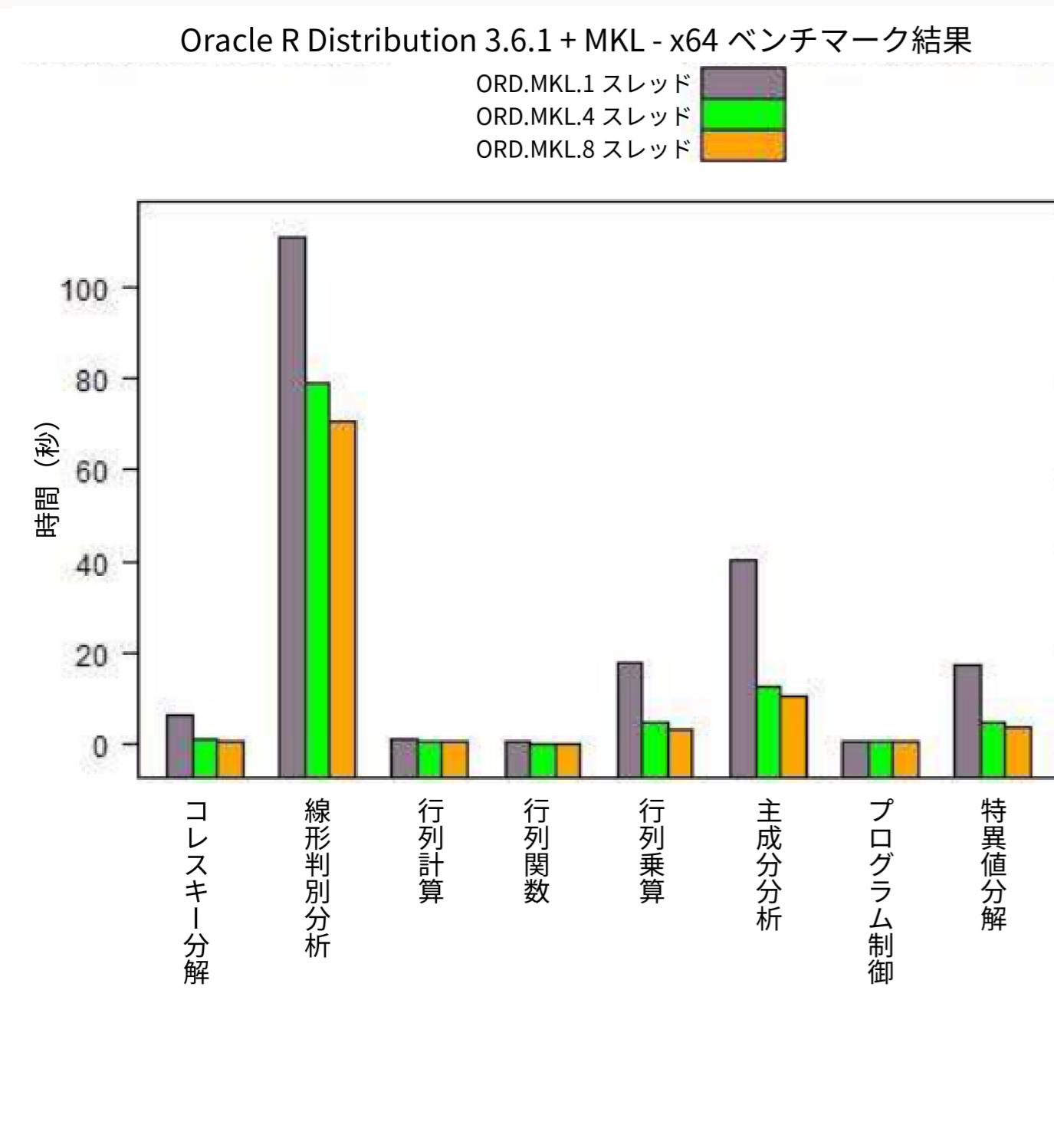
Solaris Sun Performance Library



オラクルの
サポート

- オラクルがサポートするオープンソースR（現在はR 3.6.1）の再配布パッケージ
- 動的にロードされるライブラリにより線形代数のパフォーマンスを強化
- 組込みR実行でクライアントとデータベースのパフォーマンスを向上
- Oracle Advanced Analyticsオプション、Big Data Appliance、およびOracle Linuxの顧客に対するエンタープライズ・サポート
- 無料ダウンロード
- オラクルがバグ修正とオープンソースRの拡張に寄与

MKLを使用したOracle R Distribution (ORD) のパフォーマンス



ROracleパッケージ

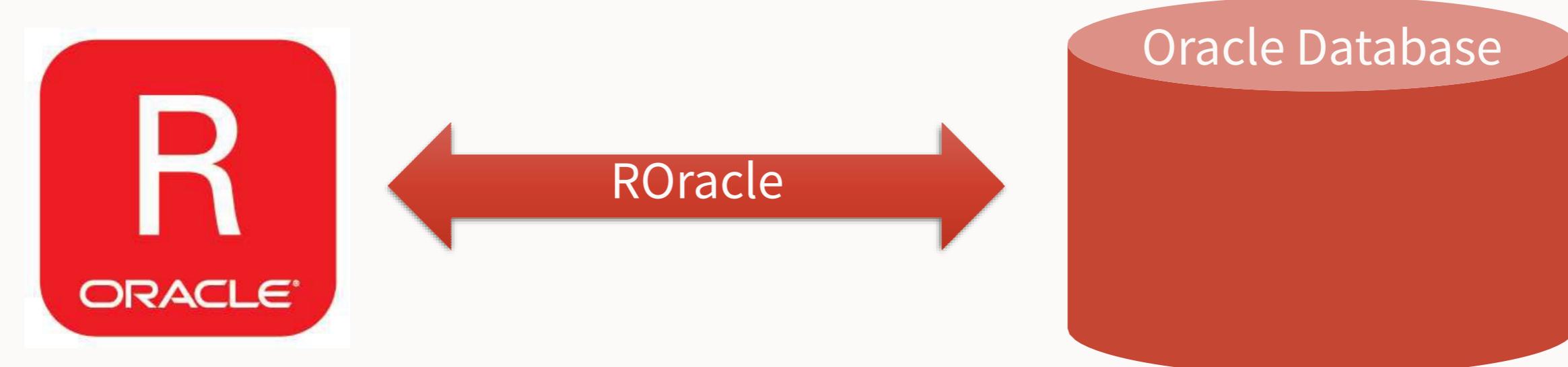
ROracle

Rパッケージにより、Oracle Databaseへのスケーラブルで高性能な接続性が実現

- CRANで一般公開されているオープンソース・パッケージ
- オラクルによって保守

R向けのOracle Database Interface (DBI)

- OCIを基盤に再実装および最適化されたドライバ
- RインターフェースからSQL文を実行
- 挿入、更新、削除のトランザクション動作が可能



トランザクション動作を可能にするROracleの例

```
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

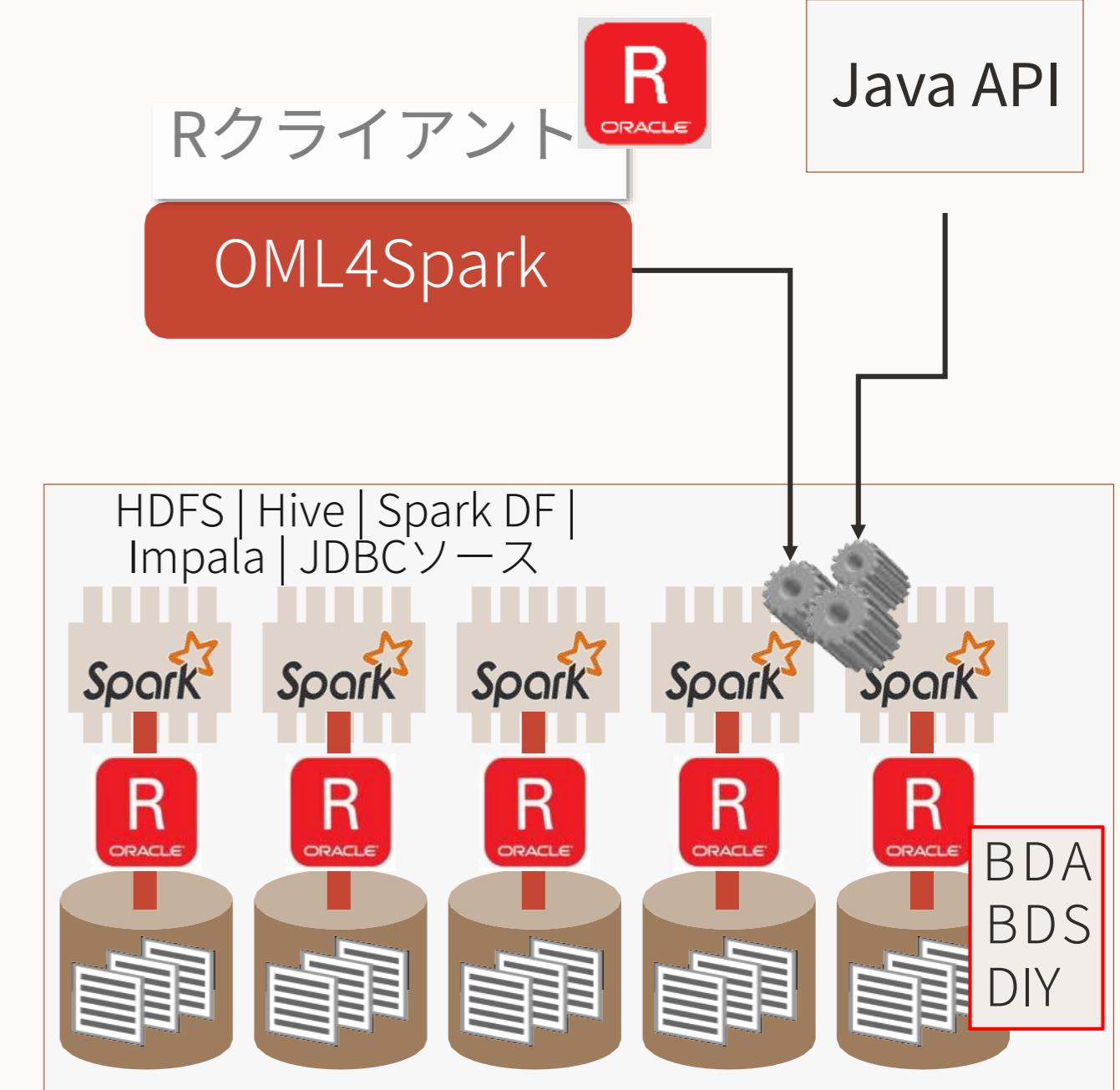
Oracle Big Data ConnectorsへのR言語APIコンポーネント

強力なデータ準備と機械学習のためにSpark 2環境を活用

幅広いデータ・レイク・ソースにわたってデータを利用

Hadoopクラスタを十分に活用してスケーラビリティと
パフォーマンスを達成

ネイティブ実装およびSpark MLlib実装による並列分散ML
アルゴリズム



Oracle Machine Learning for Spark

Oracle Big Data ConnectorsへのR言語APIコンポーネント

透過レイヤー

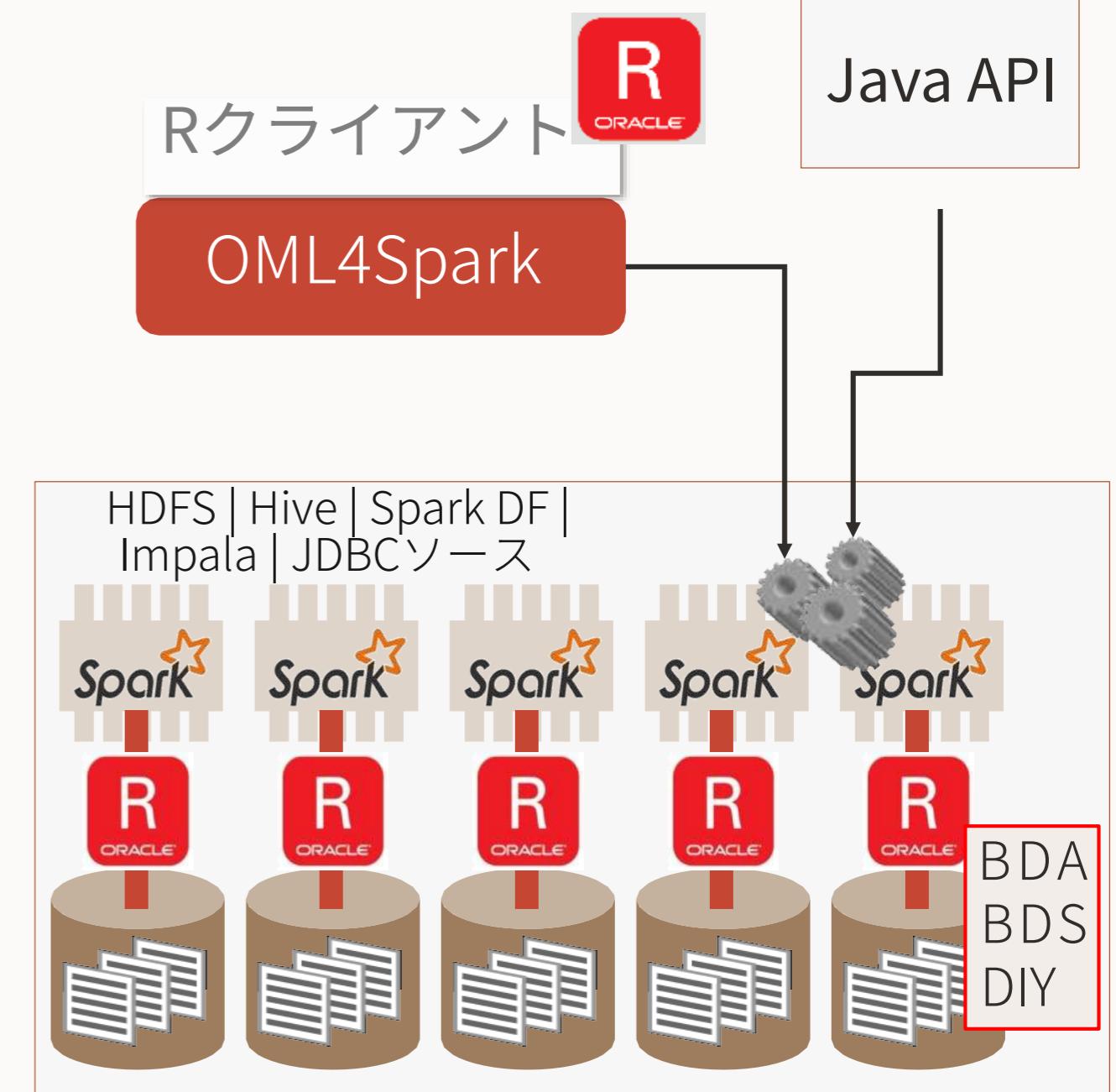
- プロキシ・オブジェクトが、ファイル・システム、HDFS、Hive、Impala、Spark DataFrame、各種JDBCソースからのデータを参照
- オーバーロードされたR関数が、機能をネイティブ言語（例：HIVE、Impalaの場合はHiveQL）に変換
- ユーザーが標準R構文でデータを操作

並列分散の機械学習アルゴリズム

- Hadoopクラスタを十分に活用することによるスケーラビリティとパフォーマンス
- SparkベースのカスタムLM、GLM、NN、k-meansのほか、Spark MLlibも利用可能
- 分かりやすいR計算式仕様を使用

カスタムRマッパー/リデューサを搭載した コンピューティング・フレームワーク

- データ並列、タスク並列の実行
- クラスタ・ノードで実行されるオープン・ソースのCRANパッケージを利用可能



OML4Sparkのパフォーマンス

ロジスティック回帰 (GLM) データがメモリ内に収まる場合

- Spark MLlibよりも最大で7倍高速

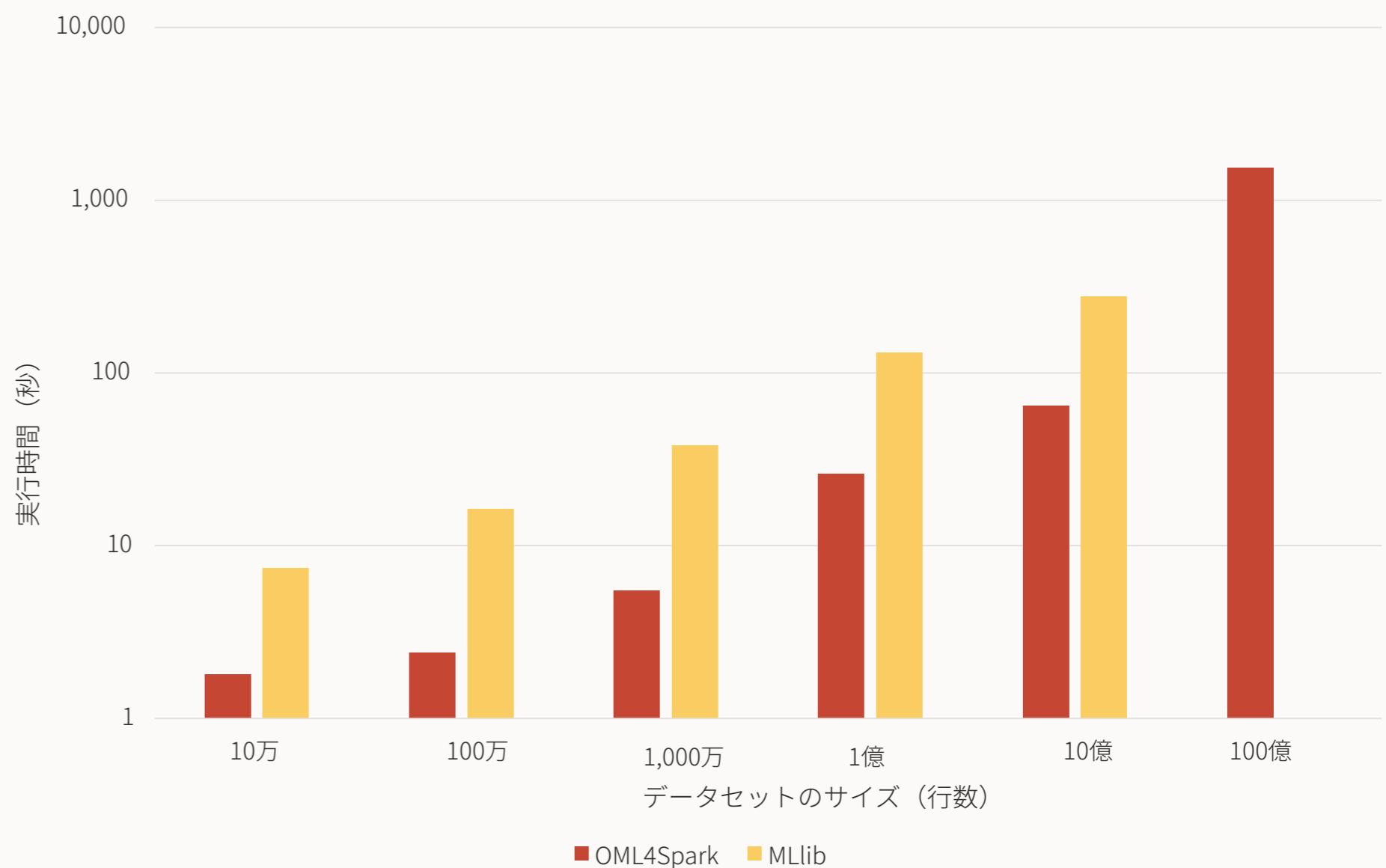
データがメモリ内に収まらない場合

- 100億行のモデルを解決可能

ベンチマーク環境

- ORAAH 2.8.0
- Big Data Appliance X7-2
- 6ノード、ノードあたり256 GBのRAM

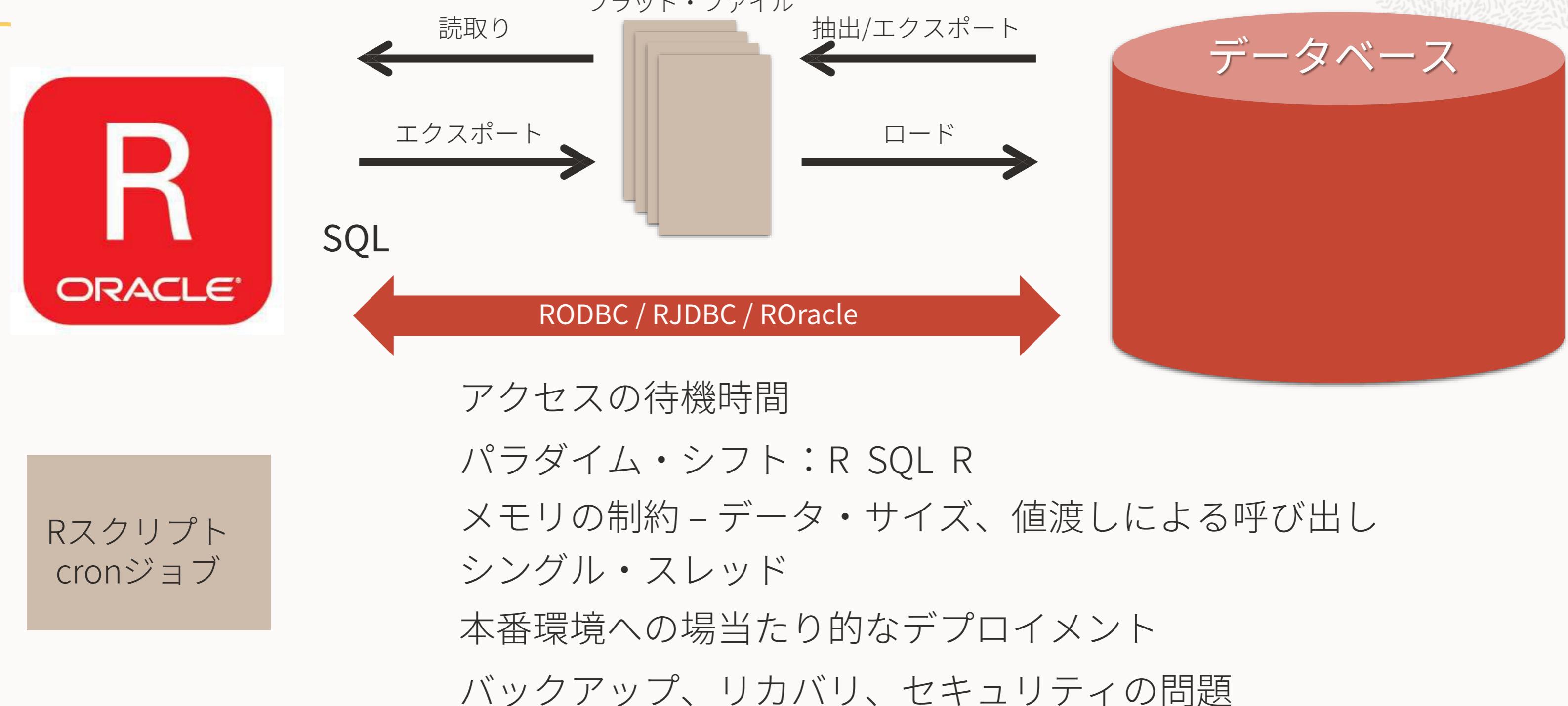
OML4SparkとSpark MLlibの比較
ロジスティック回帰 (GLM) の場合



計算式 : 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)

Rとデータベースの従来の相互作用



Oracle Machine Learning for R

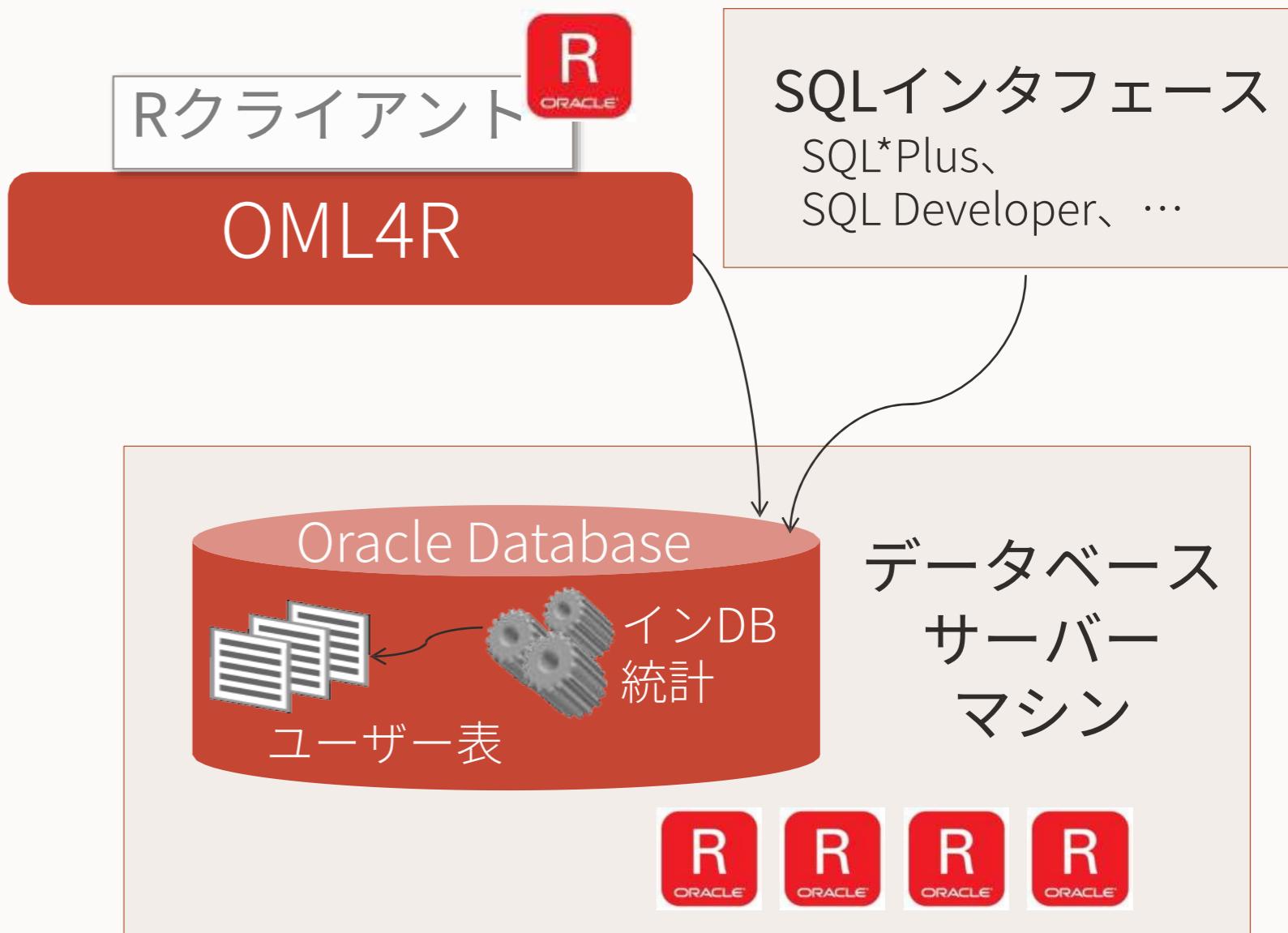
Oracle Databaseのコンポーネント

Oracle DatabaseをHPC環境として使用

並列で分散されたインデータベースの
機械学習アルゴリズムを使用

Oracle Database内で
RスクリプトとRオブジェクトを管理

Rの結果をSQL経由で
アプリケーションとダッシュボードに統合



Oracle Machine Learning for R

Oracle Databaseのコンポーネント

透過レイヤー

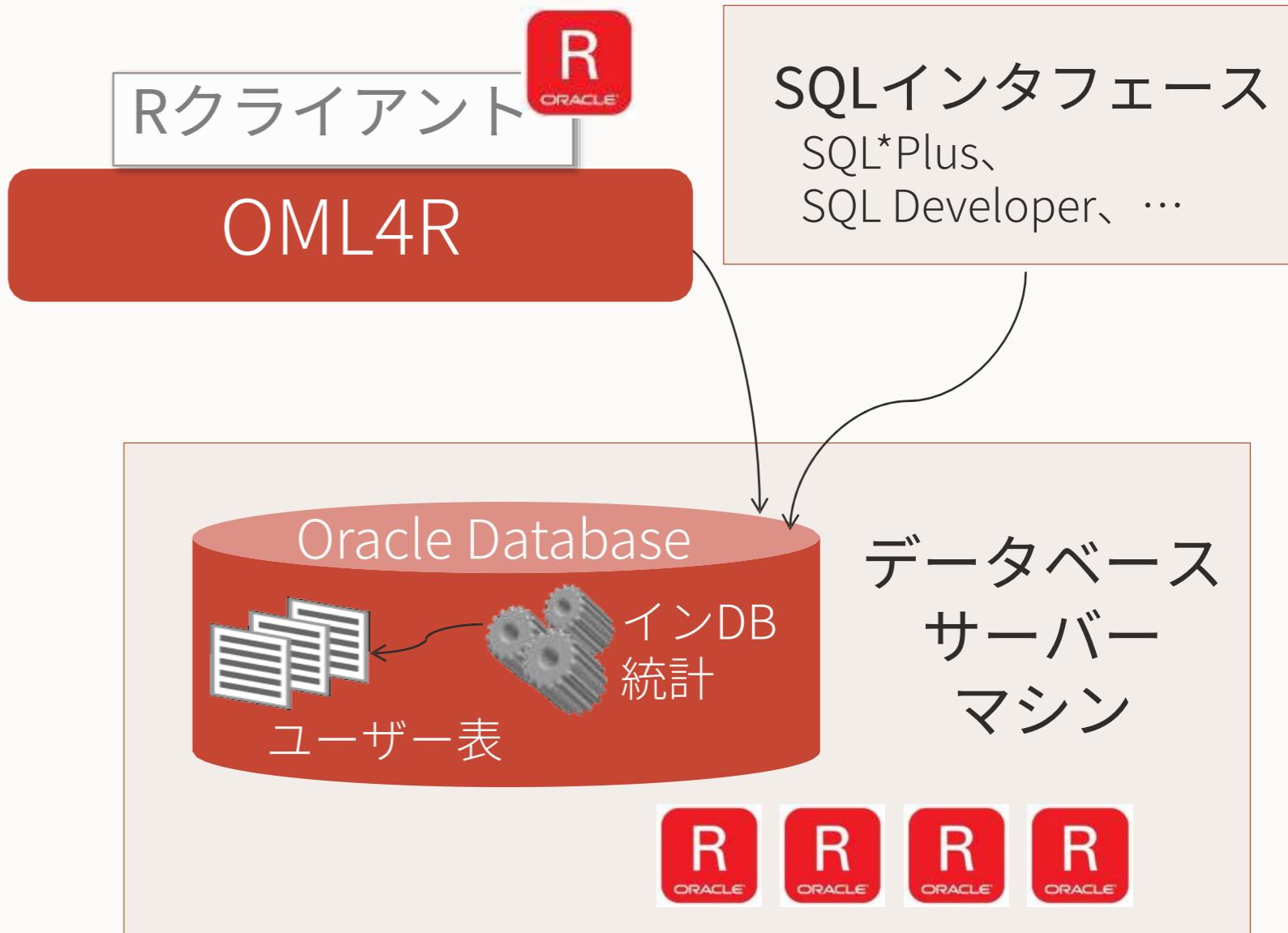
- プロキシ・オブジェクトを活用してデータベース内にデータを保持
- 機能をSQLに変換するR関数をオーバーロード
- 標準のR構文を使用してデータベース・データを操作

並列分散の機械学習アルゴリズム

- スケーラビリティとパフォーマンス
- インデータベース・アルゴリズムがOML4SQLにより公開
- データベース・サーバーで実行されるRベースの追加アルゴリズム

組込みR実行

- Oracle DatabaseでRスクリプトを管理して起動
- データ並列、タスク並列、および非パラレルの実行
- オープンソースのCRANパッケージを使用



Oracle R Enterprise (OML4R) の書籍

Amazonで購入可能

Oracle R Enterprise

Harnessing the Power of R in Oracle Database: Transform
Your Organization's Big Data Into Valuable Assets



Oracle R Enterprise Harnessing the Power of R in Oracle Database

Transform Your Organization's Big Data
Into Valuable Assets

Brendan Tierney
Oracle ACE Director

Oracle
Press®

Copyrighted Material

OML4Rアルゴリズム

+ 組込みRデータとタスクのパラレル実行を組み合わせた
アルゴリズム用のオープンソースRパッケージ

分類

- ディシジョン・ツリー
- ロジスティック回帰
- ナイーブ・ベイズ
- サポート・ベクター・マシン
- ランダムフォレスト

回帰

- 線形モデル
- 一般化線形モデル
- 多層ニューラル・ネットワーク
- 段階的線形回帰
- サポート・ベクター・マシン

クラスタ化

- 階層型k-means
- 直交パーティション化
- 期待値最大化

属性の重要度

- 最小記述長

異常検出

- 1クラス・サポート・ベクター・マシン

マーケットバスケット分析

- Apriori – 相関ルール

特徴抽出

- 非負値行列因子分解
- 主成分分析
- 特異値分解
- 明示的セマンティック分析

時系列

- 単純指数平滑法
- 二重指数平滑法

自動データ準備、パーティション化されたモデル一式、統合テキスト・マイニングをサポート

インデータベース集計関数の呼び出し

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

ソース・データは、Oracle Databaseに常駐するore.frame ONTIME_Sです

aggregate()関数は、OREフレームを受け入れるようにオーバーロードされています
aggregate()は、標準Rのdata.framesとore.framesを操作するコード間で透過的に切り替わります

ore.frameを戻します

```
R> aggdata <- aggregate(ONTIME_S$DEST,  
+                           by = list(ONTIME_S$DEST),  
+                           FUN = length)  
  
R> class(aggdata)  
[1] "ore.frame"  
attr(,"package")  
[1] "OREbase"  
R> head(aggdata)  
Group.1 x  
0 ABE 237  
1 ABI 34  
2 ABQ 1357  
3 ABY 10  
4 ACK 3  
5 ACT 33
```

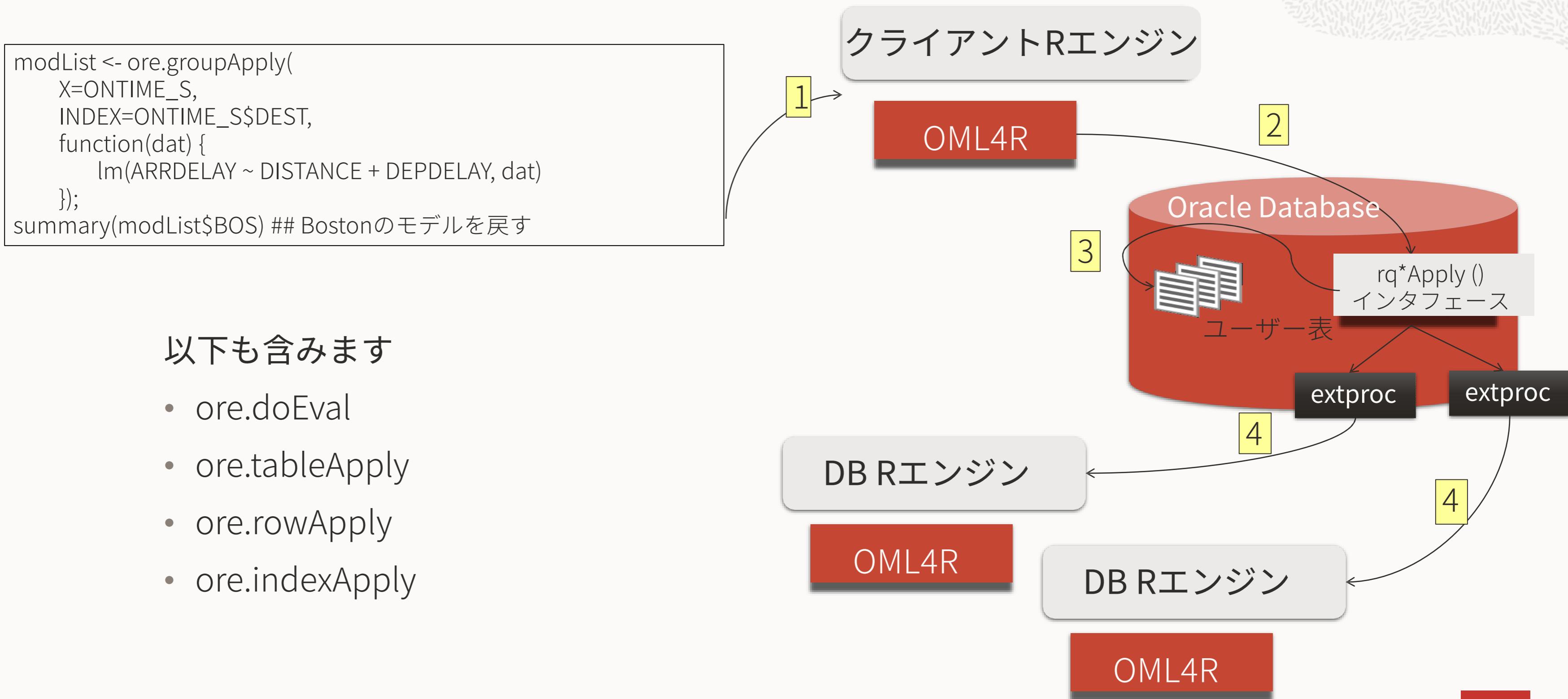


ore.groupApply – パーティション化されたデータ・フロー

```
modList <- ore.groupApply(  
  X=ONTIME_S,  
  INDEX=ONTIME_S$DEST,  
  function(dat) {  
    lm(ARRDELAY ~ DISTANCE + DEPDELAY, dat)  
  };  
  summary(modList$BOS) ## Bostonのモデルを戻す
```

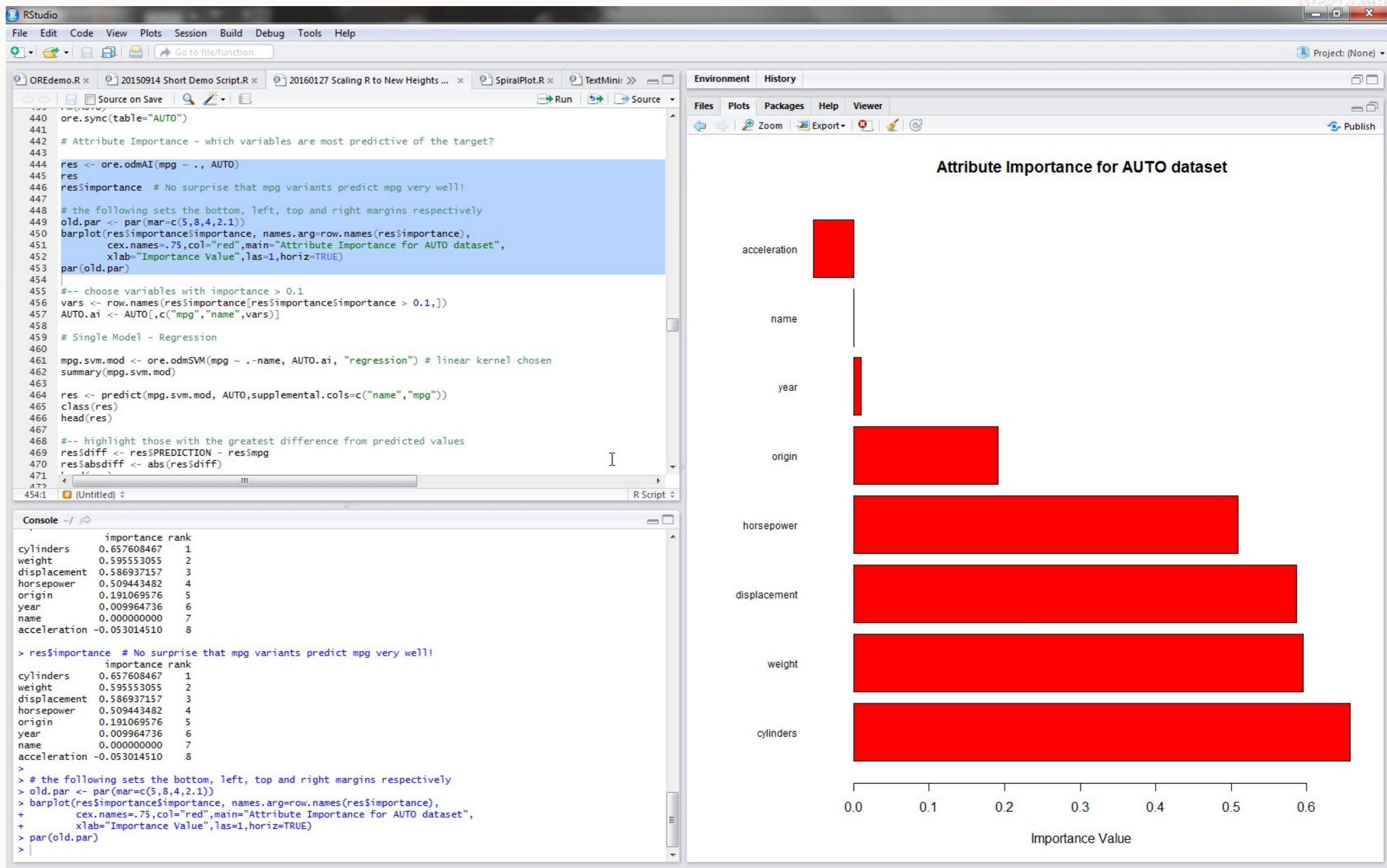
以下も含みます

- ore.doEval
- ore.tableApply
- ore.rowApply
- ore.indexApply



ore.odmAIを使用して
重要な予測変数を選択

インデータベース処理に
よりデータの移動を排除



組込みR実行 - SQLインターフェース

モデルの構築とバッチのスコアリング向け

```
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;
```

Rインターフェースを使用した統計

空間関数

- ガンマ関数
- ガンマ関数の自然対数
- ディガンマ関数
- トリガンマ関数
- エラー関数
- 補足的なエラー関数

検定

- カイ二乗検定、マクネマー検定、Bowker検定
- シンプルで重み付きのkappa検定
- コクラン-マンテル-ヘンツェル相関
- クラメールのV
- 二項検定、KS検定、t検定、F検定、Wilcox検定

Base SASと同等機能

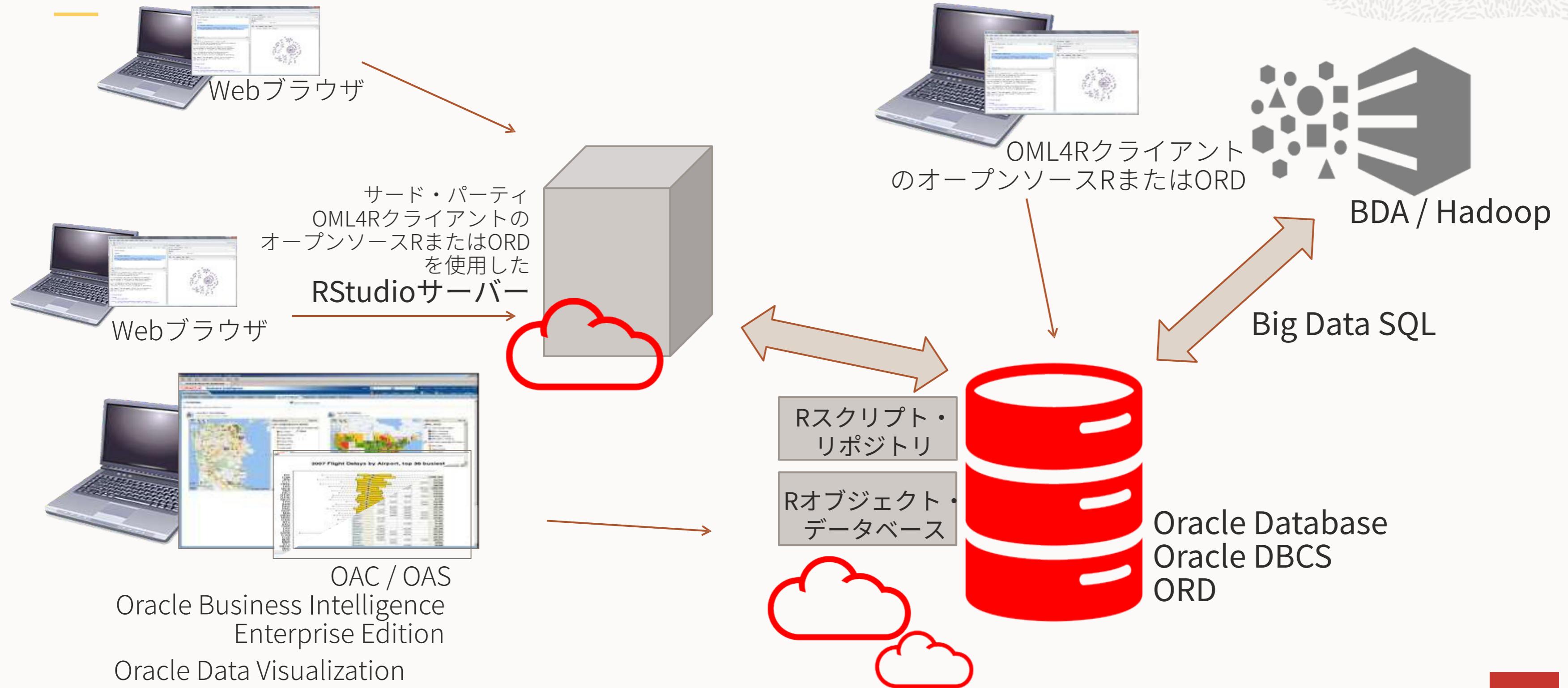
- 頻度、サマリー、ソート
- 順位、相関、一変量

密度関数、確率関数、 クォンタイル関数

- ベータ分布
- 二項分布
- コーシー分布
- カイ二乗分布
- 指数分布
- F分布
- ガンマ分布
- 幾何分布
- 対数正規分布
- ロジスティック分布

- 負の二項分布
- 正規分布
- ポアソン分布
- 符号付き順位分布
- スチューデントのt分布
- 一様分布
- ワイブル分布
- 密度関数
- 確率関数
- クォンタイル

Oracle Machine Learning for Rのデプロイメント・アーキテクチャ・オプション



まとめ

オラクルは、インデータベース機械学習でSQL、R、Python、およびノーコードUI向けのインターフェースをサポート

Rユーザーは、ビッグ・データで高度な分析を行うことが可能に

- Oracle Database
- Oracle Machine Learning for Sparkを使用したBig Data ApplianceとCloudera/Hortonworksクラスタ

オラクルのRテクノロジーによりオープンソース・ツールがエンタープライズで使用できるよう拡張

- データ分析、データ探索、機械学習
- アプリケーション開発の簡素化
- 本番環境へのデプロイメント

高いパフォーマンス、スケーラビリティ、本番環境への容易なデプロイメントを実現

追加情報

oracle.com/machine-learning

Database / Technical Details /
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.

[AskTOM OML Office Hours](#)もご覧ください

ありがとうございました

Mark Hornick
Oracle Machine Learning製品管理