19c ORACLE Database

Oracle Machine Learning

Move the Algorithms, Not the Data

ORACLE WHITE PAPER | JULY 2019

ORACLE

Disclaimer

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.

Table of Contents

"Essentially, all models are wrong, …but some are useful."

GEORGE BOX

FAMOUS TWENTIETH CENTURY STATISTICIAN

Executive Summary:
Machine Learning Algorithms Embedded in Data Management Platforms

The era of "big data" and the "cloud" is driving enterprises to change. Just to keep pace, executives and technologists must learn skills and implement new practices that leverage non-traditional data sources and technologies. As a result of sharing their "digital exhaust", customers have increased expectations of the enterprises that service them. Customers expect improved customer interactions and greater perceived value. With big data and the cloud, enterprises have the opportunity to satisfy these new expectations. Cloud, competition, big data analytics and next-generation "predictive" applications enable enterprises to achieve new goals, and deliver greater insights and better outcomes. Traditional business intelligence and analytics approaches do not deliver these insights and simply cannot satisfy the ever-growing customer expectations in this new world order created by big data and the cloud.

Unfortunately, with big data, as data volumes grow in velocity (speed of delivery), volume (size), and variety (data types), new problems emerge. Data volumes can quickly become unmanageable and immovable. Scalability, security, and information latency become new issues. In addition, dealing with unstructured text, sensor data, or spatial and graph data each introduce new analytics complexities.

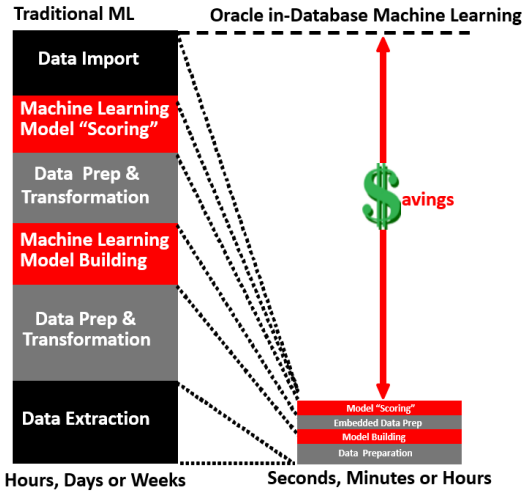
Traditional advanced analytics tools, now often referred to as *machine learning* tools, inherently have several information technology weak points: the need for data extracts and data movement, data duplication resulting in no single-source of truth, data security concerns, the need for separate and often multiple analytical tools (commercial and open source) and languages (SAS, R, SQL, Python, SPSS, etc.). These weak points become particularly egregious during the deployment phase when the worlds of data analysis and information management collide.

Traditional machine learning processes typically start with a representative sample or subset of the data that is exported to separate analytical servers and tools (SAS, R, Python, SPSS, etc.) that have been often designed for data scientists. The analytics they perform range from simple descriptive

statistical analysis to advanced, predictive, and prescriptive analytics, including machine learning. If a data scientist builds a predictive model that is determined to be useful and valuable, Information Technology (IT) organizations needs to be involved to figure out deployment. It's at that point that enterprise deployment and application integration issues become a significant challenge.

In this scenario, to make predictions on larger datasets maintained within a data warehouse, predictive models—and all their associated data preparation and transformation steps—must somehow be translated to SQL and recreated inside the database, or translated to Java for execution in an application. This model translation phase introduces tedious, time consuming and expensive manual coding steps from the original statistical language (SAS, R, and Python) into SQL and Java. DBAs and IT must somehow "productionize" these separate machine learning models within or alongside the database and/or data warehouse for distribution throughout the enterprise. This is where many advanced analytics projects fail. Some vendors will charge for specialized products and options for just the predictive model deployment capability. Add Hadoop, sensor data, tweets, and expanding big data reservoirs, and the entire data-to-actionable-insights process becomes more challenging.

Not with Oracle. Oracle delivers a big data and analytics platform that eliminates the traditional extract, move, load, prepare and analyze, export, move, import workflow. With Oracle Database 19c and Oracle Machine Learning, big data management and machine learning are combined and designed into the data management platform from the beginning. Through multiple decades of R&D investment in developing the industry's leading data management platform, Oracle SQL, Big Data SQL, Cloud SQL, Oracle Exadata, Oracle Big Data Service and integration with open source R, Oracle provides a seamlessly integrated data analytics platform in Oracle Database. Now, Oracle Database becomes a "Thinking Database."



Oracle Machine Learning Platform compared to traditional machine learning reduces the time from data to model deployment from weeks to hours.

The Oracle vision has two objectives: 1) make big data and analytics simple for a range of users involving a wide variety of data in different compute environments, and 2) make analytics results simple to deploy, whether as a service or in an application or dashboard.

Oracle Machine Learning offers a library of powerful in-database algorithms and integration with open source R that together can solve a wide variety of business problems. These algorithms are accessible via SQL, R, or user interface. Oracle Machine Learning, supported by the Oracle Advanced Analytics option to Oracle Database 19c Enterprise Edition, extends the database into an enterprise-wide analytical platform for data-driven problems such as churn prediction, customer segmentation, fraud and anomaly detection, identifying cross-sell and up-sell opportunities, market basket analysis, text mining, and sentiment analysis. Oracle Machine Learning empowers data scientists and analysts to extract knowledge, discover new insights, and make data-driven predictions—working directly with large data volumes in Oracle Database and Big Data environments.

Data scientists and analysts have choice and flexibility in how they interact with Oracle Machine Learning. Oracle Data Miner is an Oracle SQL Developer extension designed as an easy-to-use "drag and drop" workflow user interface that automates many of the steps in the machine learning process. Oracle SQL Developer is a free, integrated development environment that simplifies database development and management of Oracle Database in both traditional and Cloud deployments. When satisfied with their analytical workflows, Oracle Data Miner users can share their workflows with other analysts and/or generate SQL scripts to hand to their IT organization to accelerate solution deployment. Oracle Data Miner also provides a PL/SQL API for workflow scheduling and automation.

Oracle Machine Learning for R is supported by Oracle R Enterprise, a component of the Oracle Advanced Analytics Option to Oracle Database. OML4R makes the open source R statistical programming language and environment ready for the enterprise and big data. Designed for problems involving both large and small volumes of data, OML4R integrates R with Oracle Database. Data scientists and broader R users can take advantage of the R ecosystem on data managed by Oracle Database. R provides a rich ecosystem of software packages for data manipulation, graphics, statistical functions, and machine learning algorithms. Oracle Machine Learning for R extends R's capabilities through three primary areas: transparent access and manipulation of database data from R through overloaded R function-to-SQL translation, in-database machine learning algorithms, ease of deployment using embedded R execution.

Application developers, using the Oracle Machine Learning algorithms and Oracle Machine Learning for R / SQL can build completely automated predictive solutions that leverage the strengths of the database and the flexibly of R to integrate and deploy analytical solutions into dashboards and applications.

By integrating big data management and analytics into the same powerful Oracle Database 19c data management platform, Oracle minimizes data movement, reduces total cost of ownership and provides the fastest way to deliver enterprise-wide solutions.

Big Data and Analytics—New Opportunities and New Challenges

Gartner characterizes big data as "high volume, velocity, and/or variety information assets that demand new, innovative forms of processing for enhanced decision making, business insights or process optimization." However, for many, this is not new. Enterprises have been mining large volumes of data for years. What is new and more challenging is the increasing pace of "big data" in terms of volume, velocity and variety, which places new demands on Information Technology (IT) departments, data scientist and data analysts and the departments and lines of business they support, e.g., marketing, customer service, support, R&D, and operations.

As data volumes grow, it eventually becomes impractical to move large data volumes to separate servers for analysis. During the big data explosion, enterprises experience many problems: the cost of data movement, error and overhead introduced by data duplication, lack of data security and accountability, the evolution of data analysis "sprawl-marts", separation of data management from data analysis, and worse, increased information latency often in terms of days or weeks.

Traditional data analysis methods contribute to these problems. Data scientists and analysts typically have their own special tools that they've learned (SAS, R, SPSS, Python, Java, etc.), which require data extracts from the database, data warehouse, or data lake. This data must then be loaded to dedicated, separate hardware dedicated to data analytics. Once a data scientist builds a predictive model meeting enterprise requirements, then a new problem emerges – how to deploy that model. This typically involves deployment to applications and dashboards that support call centers, websites, ATMs, or mobile devices. The predictive model(s)—and all the associated data preparation and transformation steps—must be recreated in the destination environment to make the predictions on larger data tables. For Oracle environments, this "export → data analysis → import results" outer loop complicates data analysis unnecessarily by increasing the time and expense of the model deployment phase. IT may resort to "productionize" the models by re-implementing them using SQL inside the database.

In some database applications, models originally created using a statistical programming language need to be run as SQL functions inside the database. Having to recode models results in a time sink and errors can be introduced. For organizations who strive to be leaders, efficient data collection, data management, analysis, and deployment of predictive models are the keys to their success. Traditional data analysis methods do not suffice. Add Hadoop, sensor data, tweets, and the ever-expanding set of new data sources and the problem just gets worse.

Machine Learning

Machine learning is the process of automatically processing large volumes of data to find previously hidden patterns, discover valuable new insights and make informed predictions for data-driven problems such as:

- Predicting customer behaviors, identifying cross-selling and up-selling opportunities
- Anticipating customer churn, employee attrition and student retention
- Detecting anomalies and combating potential tax, medical or expense fraud,
- Understanding hidden customer segments and understanding customer sentiment,
- Identifying key factors that drive outcomes and delivering improved quality

Machine Learning, also sometimes referred to as predictive analytics or data mining, as a technology has been delivering measurable value for years. Predictive Analytics climbed its way up Gartner's Hype Cycle for Emerging Technologies and reached the Gartner's enviable "plateau of productivity" in 2013. Today, in 2019, machine learning is being implemented and deployed across enterprises in

solutions ranging from predicting churn and employee turnover, to flagging medical fraud and tax non-compliance. Targeted selling and real-time recommendation engines are also commonplace. As big data analytics technologies and user adoption matures and expands, predictive analytics use cases and integrated “predictive” applications that push “the art of the possible” emerge every day and constantly raise the bar for user’s expectations.

Oracle Machine Learning provides support for these data driven problems by offering a wide range of powerful machine learning algorithms implemented as SQL functions inside Oracle Database, and exposed directly through SQL or R APIs. As a result, Oracle Machine Learning algorithms leverage all related SQL features and can mine data in its original star schema representation including standard structured tables and views, transactional data and aggregations, unstructured data as found in CLOB data types (using Oracle Text to extract “tokens”), and spatial and graph data. Oracle Machine Learning in-database algorithms take advantage of database parallelism for both model building and scoring, honor security and user privilege schemes, adhere to revision control and audit tracking database features, and can mine data in their native and potentially encrypted form – inside Oracle Database.

CLASSIFICATION
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine
- Explicit Semantic Analysis

CLUSTERING
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

ANOMALY DETECTION
- One-Class SVM

TIME SERIES
- Forecasting - Exponential Smoothing
- Includes popular models e.g. Holt-Winters with trends, seasonality, irregularity, missing data

REGRESSION
- Linear Model
- Generalized Linear Model
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- LASSO

ATTRIBUTE IMPORTANCE
- Minimum Description Length
- Principal Comp Analysis (PCA)
- Unsupervised Pair-wise KL Div
- CUR decomposition for row & AI

ASSOCIATION RULES
- A priori/ market basket

PREDICTIVE QUERIES
- Predict, cluster, detect, features

SQL ANALYTICS
- SQL Windows
- SQL Patterns
- SQL Aggregates

FEATURE EXTRACTION
- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

TEXT MINING SUPPORT
- Algorithms support text
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA) for document similarity

STATISTICAL FUNCTIONS
- Basic statistics: min, max, median, stdev, t-test, F-test, Pearson’s, Chi-Sq, ANOVA, etc.

R AND PYTHON PACKAGES
- Third-party R and Python Packages through Embedded Execution
- Spark MLlib algorithm integration

Oracle Machine Learning provides a wide variety of algorithms and analytics functionality.

Move the Algorithms, Not the Data

Data is big; algorithms are small. Hence, it makes logical sense to move the algorithms to the data rather than move the data to the algorithms. Oracle realized this big data and analytics data challenge

in 1999 when it acquired Thinking Machines Corporation's data mining technology and development team. At that time, Oracle commenced on a strategy to develop traditional and innovative machine learning algorithms and statistical functions as native SQL functions with full SQL language support. With Oracle Machine Learning, machine learning algorithms run as native SQL functions – not as PL/SQL scripts, call-outs, or extensibility framework add-ins. Models are first class database objects that can be built, applied, shared, and audited.

In the early 2000's, starting in Oracle Data Mining Release 9.2, Oracle's first data mining algorithms took advantage of available core Oracle Database strengths—specifically, counting, parallelism, scalability and other database architectural underpinnings. Essentially, the first two Oracle data mining algorithms, Naïve Bayes and A Priori, are based on counting principles. They count everything very quickly and then assemble conditional probability predictive models—all 100% inside the database. Neither the data, the predictive models, nor the results ever left the database.

This Naïve Bayes algorithm can quickly build predictive models to predict e.g., "Who will churn?", "Which customers are most likely to purchase Product A?", or "What is the probability that an item will fail?" Consider the following example. Say we are interested in selling Product A (e.g., a motorcycle or $500 shoes). The Naïve Bayes algorithm considers all the customers who purchased Product A, and counts how many customers were male vs. female, identifies how many rent an apartment versus owns their own home, or how many have children (and how many children). Each of these involves counts that, taken together, can form a complex conditional probability model that accurately predicts whom we should target to increase our likelihood of selling more of Product A.

The A Priori "market basket analysis" algorithm counts items in each customer's transactional "basket" while looking for co-occurring items. E.g., the algorithm may determine that A and B frequently appear together, and then provide conditional probability rules, such as:

> IF "Cereal" AND "Bananas" appear in the same customer's basket,
> THEN the "Milk" is also likely to appear in the basket.
> WITH Confidence = 87%, and Support = 11%.

Armed with these types of new customer insights from Oracle Machine Learning, a store could decide to place the milk near the cereal and bananas, or put it in opposite sides of the store (to encourage customers to walk more in the store and add more items to their baskets), offer new promotional "breakfast kit" product bundles, or make real-time customer specific recommendations as the customer

checks out. This is just a simple example of the ways that big data analytics can find “actionable insights” from data. Obviously, more data, scalable machine learning, and fast enterprise-wide deployment can open new doors to many big data and analytics application and solution possibilities.

SQL and R Support

Where SQL has been the standard language for data management for 40+ years, for data analysis, various languages and tools compete—R, Python, SAS, SPSS, Matlab and others. These have been long time favorites, but in recent years, open source R and Python have surged to the top of the pack. Per the KD Nuggets data mining industry community annual polls (http://www.kdnuggets.com/polls/), R and SQL currently compete for #1 and #2 positions, respectively.

The good news is that Oracle Machine Learning supports multiple languages—SQL and R (and soon, Python). There are legions of developers who know SQL for data management and Oracle provides support for data mining and machine learning via Oracle Machine Learning for SQL and provides tight, industry leading integration with open source R via Oracle Machine Learning for R.

Most Oracle customers are familiar with SQL as a language for query, reporting, and analysis of structured data. It is the de facto standard for data processing and supports most BI tools. R is a widely popular and powerful open source programming language and statistical environment, which is free and, like Python, taught in most data science educational programs. A growing number of data analysts, data scientists, researchers, and academics start by learning to use R, leading to a growing pool of R programmers who can now work with their data inside Oracle Database using either SQL or R.

In-Database Processing with Oracle Machine Learning

Oracle Machine Learning extends Oracle Database into an advanced analytics platform for big data analytics—essentially a “Thinking Database.” With Oracle, powerful analytics are performed directly on data in the database. Results, insights, and real-time predictive models are available and managed by the database.

An Oracle Machine Learning model is a database schema object, built by invoking a PL/SQL API or corresponding OML4R function that prepares the data and discovers the hidden patterns, which can then be used to score data via built-in machine learning SQL functions. When building models, Oracle Machine Learning leverages standard scalable database technology (e.g., parallel execution, bitmap indexes, aggregation techniques) as well as custom-built technologies (e.g., recursion within the

parallel infrastructure, IEEE float, automatic data preparation for binning, handling missing values, support for unstructured data i.e. text).

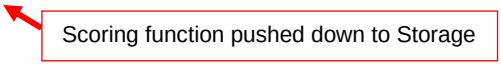
To highlight the benefit of in-database model building, one benchmark involving 640 million rows of the ONTIME airline data set on an M8-2 Exadata hardware showed SVM with the SGD solver completed in 83 seconds on 900 attributes, while GLM Regression completed in 59 seconds. This was able to leverage 512 degrees of parallelism – significantly taking advantage of the M8-2 hardware.

The power of in-database algorithms as SQL functions is even more evident when scoring machine learning models, whether in batch of in support online transactional processing (OLTP) environments. Scoring data with these models is amazingly fast – millions of records in seconds – since it amounts to a row-wise function.

Using the "smart scan" technology of Oracle Exadata, SQL predicates and OML predictive models get pushed down to the storage layer for execution. In both cases, only those records that satisfy the predicates are pulled from disk for further processing inside the database.

For example, find the US customers likely to churn (more than 80% probability based on the model called "churn_model"):

```
select cust_id from customers where region='US'
   and prediction_probability(churn_model,'Y' using *) > 0.8 ;
```



Automatic Data Preparation, Data Types, Star Schemas and "Nested Tables"

When analyzing data, analysts often make explicit decisions about how to bin data, deal with missing values, and often reduce the number of variables (feature selection) to be used for model building. Oracle Machine Learning automates most of the steps typically required in machine learning projects. Today, Automatic Data Preparation (ADP) automatically bins numeric attributes using default and user-customizable binning strategies, such as equal width, equal count, and user-defined bins. Similarly, ADP bins categorical attributes into N top values and "other" or user-defined bins. ADP will automatically replace missing values by a statistical value such as mean, median, or mode. This avoids simply removing records with missing values from the analysis or requiring the user to performing missing value replacement manually. The same statistics gathered and used by ADP when

preparing data for model building are used when applying the models to new data. Users can of course override ADP settings.

Oracle Machine Learning provides support for feature selection methodology called Attribute Importance using either the Minimum Description Length algorithm or the CUR matrix decomposition method. It also supports feature extraction through Principal Components Analysis, Non-Negative Matrix Factorization, Explicit Semantic Analysis and Singular Value Decomposition. However, each Oracle Machine Learning algorithm has its own built-in automated strategies for attribute reduction and selection so an explicit feature reduction step is optional. Users can control algorithm and data preparation settings explicitly or accept the intelligent defaults.

Transactional data, which includes purchases, transactions, and events, is often important to build good predictive models. Oracle Machine Learning analyses this data in its native transactional form by leveraging Oracle Database aggregation functions to summarize and then join summary vector data (e.g., item purchases) to other customer data to provide a wider “360 degree” customer view. Oracle Machine Learning algorithms ingest this aggregated transactional attribute as a “nested table.” The Oracle Machine Learning algorithms process records as triplets: Unique_ID, Attribute_name, and Attribute_value. That’s just part of the secret sauce of how Oracle Machine Learning leverages the core strengths of Oracle Database. The market basket analysis algorithm can analyze this data in its native transactional data form to find co-occurring items in baskets.

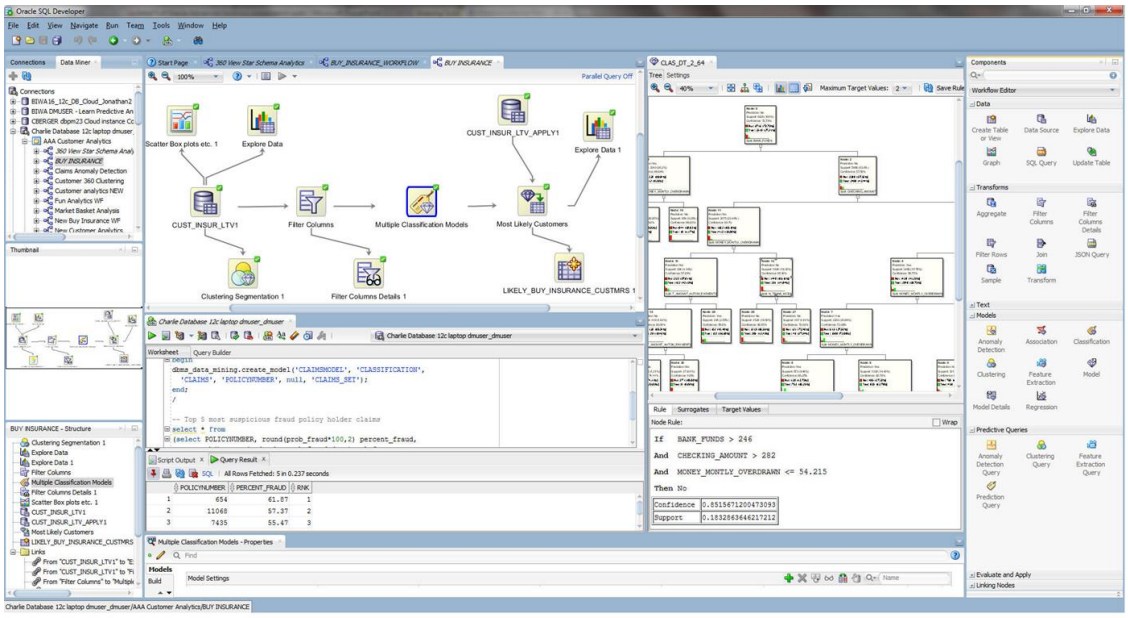
Unstructured text data is processed in a similar automated fashion using the multi-language support of Oracle Text to “tokenize” CLOB data type columns containing, e.g., raw text, Word, and Adobe Acrobat documents. Oracle Machine Learning uses Oracle Text – a free feature in Oracle Database – to pre-process text into vectors of words and word coefficients (TF-IDF—term frequency-inverse document frequency). Oracle Machine Learning algorithms treat the unstructured attributes as additional input attributes, e.g., police comments, physician’s notes, resumes, emails, articles, abstracts, that are joined with traditional structured data, e.g., age, income, occupation. Spatial data, web clicks, and other data types can also be joined for use for model building and scoring.

Oracle Data Miner; a SQL Developer Extension

Oracle Data Miner, an extension to Oracle SQL Developer, is designed for “citizen data scientists” who may prefer an easy to use UI and don’t necessarily want to or know how to program in either SQL or R. Oracle Data Miner enables data analysts, business analysts, and data scientists to work directly with data inside the database using Oracle Data Miner’s “drag and drop” workflow paradigm. Oracle

Data Miner workflows capture and document the user's analytical methodology. Users can save and share their workflows with others to automate and publish their advanced analytics methodologies.

Data analysts easily learn how to use Oracle Data Miner and can quickly visualize and explore the data graphically, prepare and transform their data, build and evaluate multiple machine learning models, and use extensive model viewing and model evaluation viewers. Then, they can apply Oracle Machine Learning models to new data or generate SQL and PL/SQL scripts to deploy their analytical workflow. These scripts can be passed to DBAs for immediate deployment within Oracle Database. Application developers can programmatically execute workflows using the Oracle Data Miner PL/SQL workflow API. This enables easily integrating predictive methodologies into applications for wider use throughout the enterprise.



Oracle Data Miner, a SQL Developer extension, provides a drag and drop workflow user interface to explore data, build, evaluate and apply predictive models, and deploy advanced analytics methodologies as SQL and PL/SQL scripts.

Oracle Data Miner supports both simple and complex advanced analytical methodologies. For example, users may want to combine transactional data, demographic data, customer service data, and customer comments to assemble a 360-degree customer view. They may decide to perform clustering on the customers to pre-assign them to customer segments, and then, for each segment, build separate classification, regression or anomaly detection models for better accuracy.

Oracle Machine Learning for R—Integrating Open Source R with Oracle Database

Oracle Machine Learning for R, supported by Oracle R Enterprise, makes the open source R statistical programming language and environment ready for the enterprise and big data. “R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering,) and graphical techniques, and is highly extensible.” (See https://www.r-project.org/) R’s strengths include being open source, powerful and extensible, and having an extensive array of graphics and statistical packages that the R user community regularly grows and enhances. R’s challenges are that it is memory constrained, single threaded, and not generally considered to be “industrial strength.”

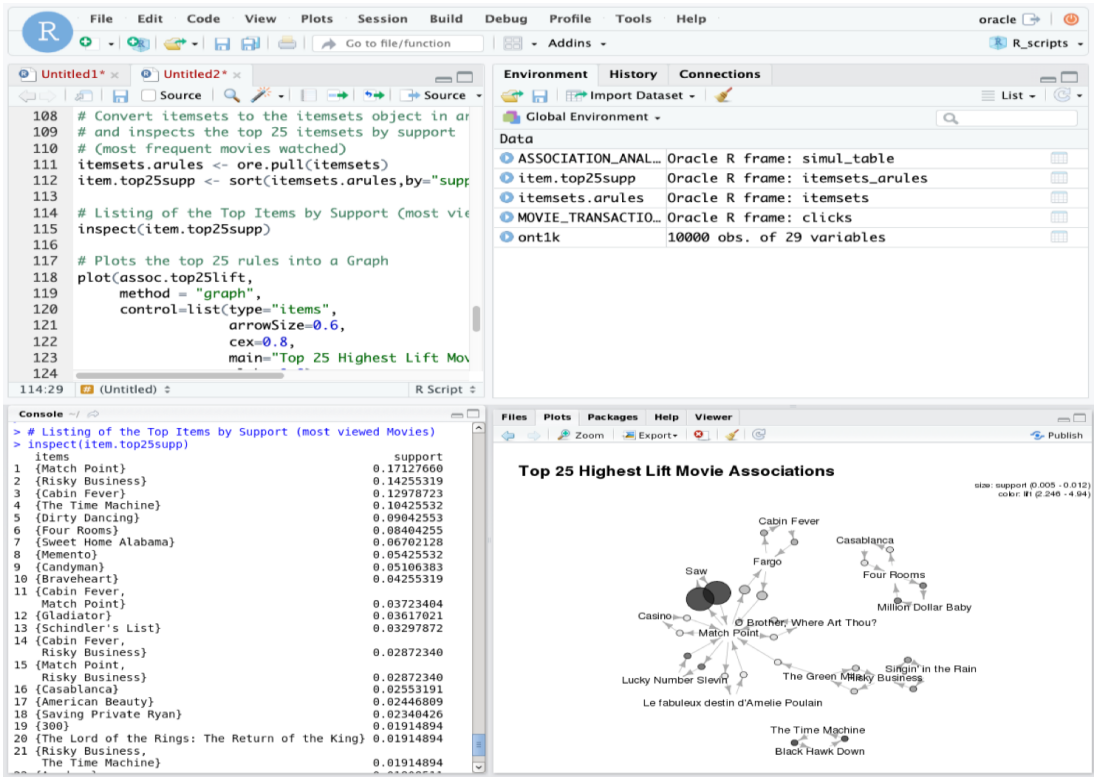
Oracle Machine Learning for R integrates R with Oracle Database, maps R functions to equivalent SQL and in-database algorithms. A combination of R packages – provided through Oracle R Enterprise – and Oracle Database features enable R users to operate on database-resident data without using SQL and to execute R user-defined functions in one or more embedded R engines that run on the database server machine. Data scientists and analysts can develop, refine, and deploy R code that leverages database parallelism and scalability and in-database algorithms without having to learn SQL.

Oracle Machine Learning for R overloads a wide range of R functions, referred to as the *transparency layer*, that transparently convert R functionality into SQL for in-database execution. With these, R programmers can create R objects that access, analyze, and manipulate data that resides in the database. The database automatically optimizes the SQL code to improve corresponding query efficiency.

With embedded R execution, users can define user-defined functions stored in the R script repository for execution at the database server machine. Oracle Database manages the spawning of R engines and loading of data to those R engines. Embedded R execution allows developers to extend Oracle Machine Learning’s native functionality by creating user-defined function that invoke functionality from open source R packages, such as those from CRAN. These user-defined R functions can be invoked in a non-parallel, data-parallel, or task-parallel manner from both R or SQL.

Users, who prefer to work in R to analyze their data, may use any R IDE, such as RStudio, to connect to an Oracle Database instance and use Oracle Machine Learning for R. Upon establishing a connection, users can obtain proxy data.frame objects for the tables and views is their scheme. These

proxy objects, of type ore.frame, can then be used in transparency layer functions, in-database machine learning algorithms, and embedded R execution.



Oracle Machine Learning for R invoking in-database Oracle Machine Learning algorithms (e.g., Apriori Association Rules) from an RStudio Server UI.

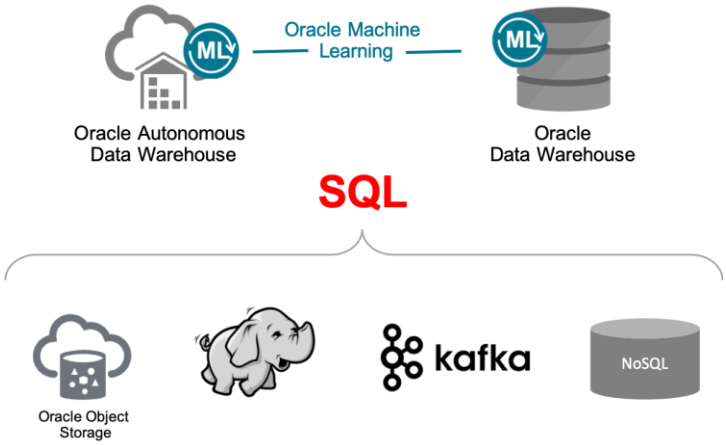
Oracle Machine Learning for Spark

Big data is often stored in Hadoop clusters, on so-called *data reservoirs* or *data lakes*. The data environment outside Oracle Database introduces new data management and data analysis challenges. Cloud SQL addresses this challenge by extending SQL processing to Hadoop via Oracle Big Data Service. Using “smart scan” technology developed for Exadata, Cloud SQL pushes down SQL logic to operate on Hive tables directly where the underlying data reside. Data analysts can now more easily take advantage of new big data sources containing data of possibly unknown value, and combine them with data of known value managed inside a database or data warehouse.

Data stored in Hadoop, however, may be voluminous and have a sparse representation (transactional format). Given that much of the data may come from sensors, Internet of Things, “tweets” and other

high volume sources, users can leverage Cloud SQL to aggregate data at various levels, e.g., counts, maximum values, minimum values, thresholds counts, averages, sliding SQL window averages.

In this case, one of the options available from Oracle Machine Learning is to filter “big data” by reducing it directly in the data lake, join it to other database data using Cloud SQL and then mine *everything* inside the Oracle Database using Oracle Machine Learning.



SQL and Cloud SQL enable data analysts to access, summarize, filter and aggregate data from both Hadoop servers and the Database and combine them for a more complete 360-degree customer view and build predictive models using Oracle Machine Learning.

Another option is to mine data directly against the big data cluster, by using the Oracle Machine Learning for Spark, which is supported by Oracle R Advanced Analytics for Hadoop, a component of the Big Data Connectors. OML4Spark R API provides functions for manipulating data stored in a local File System, HDFS, HIVE, Spark DataFrames, Impala, Oracle Database, and other JDBC sources. OML4Spark takes advantage of all the nodes of a Hadoop cluster for scalable, high performance machine learning in a Big Data environment. OML4Spark functions use the expressive R formula object optimized for Spark parallel execution.

OML4Spark brings custom LM, GLM and MLP Feed Forward Neural Networks algorithms that run on top of the Spark infrastructure. While these algorithms scale better and run faster than the open-source alternatives of Apache SparkML algorithms, OML4Spark provides interfaces to SparkML algorithms as well. R functions wrap SparkML algorithms within the ORAAH framework using the R formula specification and Distributed Model Matrix data structure. ORAAH's SparkML R functions can be executed either on a Hadoop cluster using YARN (to dynamically form a Spark cluster), or on a dedicated standalone Spark cluster.

Cloud SQL and OML4Spark can be combined from Oracle Database or Autonomous Database to address large, complex data-driven problems where the source data and patterns to be discovered may lie in big data, relational data, or some combination of the two. OML4Spark provides options for machine learning processing outside Oracle Database or as a powerful component of larger, complex machine learning pipelines.

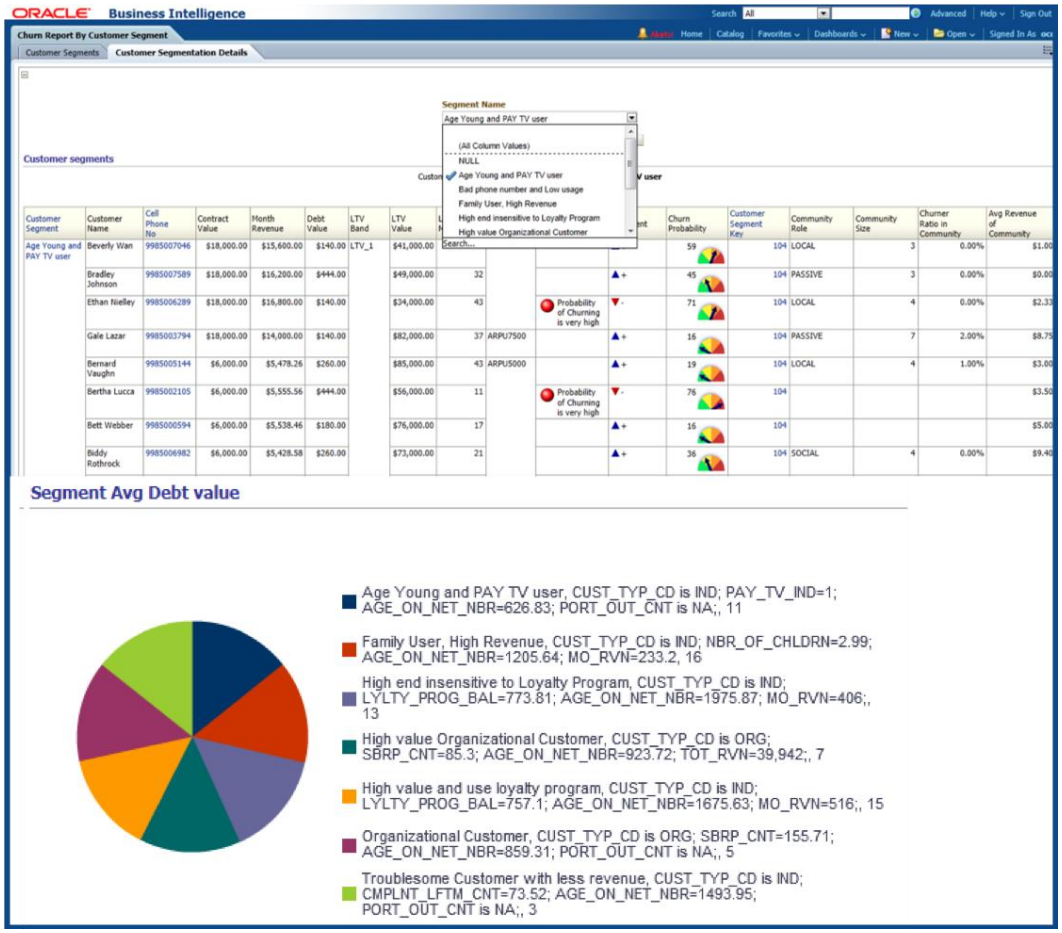
A Platform for Developing Enterprise-wide Predictive Analytics Applications

Oracle's strategy of making big data and big data analytics simple makes it easier to develop, refine and deploy predictive analytics applications—all is part of database functions. The data, user access, security and encryption, scalability, applications development environment and powerful advanced analytics are available in the data management and data analytics platform—Oracle Database. Now, it is easy to add predictive and real-time actionable insights into any enterprise application, BI dashboard, or tool that can speak SQL to the Oracle Database.

Oracle has been developing predictive analytics applications for over a decade. Oracle provides next-generation predictive applications on premise and in the cloud, including:

- Oracle Human Capital Management Predictive Workforce
- Oracle Adaptive Intelligence for Manufacturing
- Oracle Content and Experience
- Oracle Integration Cloud
- Oracle Customer Relationship Management Sales Prediction Engine
- Oracle Adaptive Access Management's Identity Management
- Oracle Retail Customer Analytics
- Oracle Predictive Incident Monitoring Premium Service
- Oracle Communications Industry Data Model
- Oracle Retail Industry Data Model
- Oracle Airlines Industry Data Model
- Oracle Utilities Industry Data Model
- Oracle Depot Repair

Oracle HCM Predictive Workforce application delivers pre-built Oracle Machine Learning predictive analytics for employee attrition, employee performance and "What if?" analysis.



Oracle Communications Data Model embeds pre-built predictive models for churn prediction, customer profiling, identifying churn factors, cross-sell, customer lifetime value (LTV) and customer sentiment.

Conclusion

Traditional BI and analytic approaches simply cannot keep pace with requirements in the era of “big data” and “cloud.” For organizations that strive to be leaders leveraging these new technologies, the prompt capture and collection of data of known and unknown value, the proper data management, assembly of relevant data, analysis, automation, and deployment of the actionable insights are the keys to success.

Oracle Machine Learning collapses the traditional extract, move, load, analyze, export, move, import paradigm all too common today. Oracle Machine Learning delivers a library of scalable, parallelized, machine learning algorithms within the Oracle Database 19c and a Big Data cluster. These libraries provide machine learning techniques including clustering, regression, classification, association rules, feature extraction, time series, and anomaly detection. Oracle Machine Learning exposes these algorithms as SQL functions accessible via SQL (Oracle Machine Learning for SQL), the Oracle Data Miner “drag and drop” workflow UI, through tight integration with open source R (Oracle Machine Learning for R) and inside a big data environment (Oracle Machine Learning for Spark).

Because Oracle Machine Learning machine learning algorithms execute in the Oracle Database, they take full advantage of Oracle Database scalability, security, integration, cloud, structured and unstructured data mining capabilities. This make Oracle the ideal platform for big data and analytics solutions and applications either on-premises or on Oracle Cloud. Oracle's multiple decades of leading edge data management experience is harnessed and combined with the strategy of "moving the algorithms to the data" and avoiding "moving the data to the algorithms".

By integrating big data management and analytics into a single unified Oracle platform, Oracle reduces total cost of ownership, eliminates data movement, and delivers the fastest way to deliver enterprise-wide predictive analytics solutions and applications.

ORACLE

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US


blogs.oracle.com/oracle


facebook.com/oracle


twitter.com/oracle


oracle.com

Hardware and Software, Engineered to Work Together

Oracle Machine Learning white paper
July 2019


Oracle is committed to developing practices and products that help protect the environment