Oracle Machine Learning
Technical Brief

Move the Algorithms, Not the Data

April 2021, Version 1.0
Copyright © 2021, Oracle and/or its affiliates
Public
Purpose statement

This document provides an overview of features and enhancements included in Oracle Machine Learning. It is intended solely to help you assess the business benefits of leveraging the Oracle data management and analytics platform with Oracle Machine Learning to plan your data-driven data science, machine learning, and information technology projects.

Intended Audience

This technical brief is intended for executives, LOB managers, and practitioners who are looking to provide scalable machine learning capabilities to the data scientists, data analysts, data engineers, and application and dashboard developers across their enterprise.

Disclaimer

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. 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 in this document remains at the sole discretion of Oracle. Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.
Table of contents

Purpose statement 2
Intended Audience 2
Disclaimer 2
Introduction 4
Executive Summary 7
What is Machine Learning? 11
Move the Algorithms, Not the Data 11
What is Oracle Machine Learning? 11
Support for SQL and open source R and Python 12
In-Database Processing with Oracle Machine Learning 12
Automation drives productivity 13
Integrating Open Source R and Python with Oracle Database 13
Oracle Machine Learning AutoML User Interface 15
Oracle Machine Learning Services 15
Oracle Data Miner—a SQL Developer Extension 16
Oracle Machine Learning in Data Lakes 17
   “Boil down” the data 17
   Oracle Machine Learning for Spark 18
Oracle Machine Learning—the Fastest Way to Deliver AI Applications 19
Conclusion 19
For more information 20
**Introduction**

Oracle delivers the data management platform that enables data science teams to analyze data where it resides at scale and with minimal data movement, whether in the database or data lake. Oracle Machine Learning is a key part of Oracle’s converged database strategy, which combines the functionality of multiple special-purpose databases while operating faster and with greater reliability. With a converged database and autonomous database self-service tools, users reduce complexity and management for an end-to-end optimized experience.

Oracle Machine Learning empowers data scientists and data 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. Oracle Machine Learning empowers application developers by making it easy to deploy and use machine learning models and solution in applications and dashboards. In this technical brief, we introduce Oracle Machine Learning and how it enables enterprise data science teams to achieve greater value from data.

Users have choice and flexibility in how they interact with Oracle Machine Learning using popular languages such as SQL, Python and R, or using interfaces such as Oracle Machine Learning Notebooks, Oracle Data Miner user interface, and Oracle Machine Learning AutoML User Interface. In addition, OML Services supports model management and deployment with Autonomous Database from a REST interface.

On Oracle Autonomous Database, users have Oracle Machine Learning Notebooks—based on Apache Zeppelin technology—that can run SQL, PL/SQL and Python scripts, and use markdown for in-notebook documentation. OML Notebooks extends Oracle Autonomous Database as a platform for machine learning and data science, supporting data scientists, data analysts, application developers, and IT professionals. OML Notebooks facilitates collaboration through real-time concurrent editing sessions, integrated graphics, easy sharing of notebooks and templates, as well as permission-based access, versioning, and job scheduling, while maintaining the security and enabling permissions inherited by the data from Oracle Database.

On Oracle Autonomous Database, Oracle Machine Learning AutoML User Interface is a productivity tool for data scientists and to enable non-expert users to benefit from powerful in-database algorithms. Users build models with minimal input—the data and the target to be predicted—and the tool does the rest to produce models guided by machine learning meta-models. OML AutoML UI also supports model deployment to OML Services, along with cognitive text analytics.

Oracle Machine Learning Services provides a common framework for machine learning model management and deployment with Oracle Autonomous Database. It supports application development with a REST interface, which extends the
model deployment through SQL, Python, and R discussed next. These model management and deployment services enable users to deploy both in-database machine learning model as well as models exported in ONNX format, which may have been produced separately from packages such as Tensorflow, PyTorch, among others. In addition, OML Services supports cognitive text analytics, like extracting topics and keywords, sentiment analysis, and text summary and similarity.

Oracle Machine Learning for SQL (OML4SQL) offers powerful in-database machine learning algorithms for model building using PL/SQL and scoring data using SQL queries to make predictions. These same in-database algorithms are also exposed through Python and R interfaces, Oracle Machine Learning Notebooks, and Oracle Data Miner. This extends the database as an enterprise-wide analytical platform for data-driven problems such as churn prediction, customer segmentation, fraud and anomaly detection, cross-sell and up-sell opportunity identification, market basket analysis, and predictive maintenance, and many others.

Oracle Machine Learning for Python (OML4Py) enables data scientists and Python users to take advantage of the Python environment on data managed by Oracle Database and Oracle Autonomous Database. Python provides a suite of software packages for data manipulation, data exploration, graphics, statistical functions, and machine learning algorithms. OML4Py extends Python's capabilities through four primary areas: (i) exploration and preparation of database data using familiar Python syntax and functions, (ii) access to in-database machine learning algorithms using a natural Python interface, (iii) ease of deployment of user-defined Python functions with embedded Python execution through a REST API, and (iv) automated machine learning (AutoML). For transparent access, OML4Py overloads select Python functions, such as those on Pandas DataFrame objects, and produces SQL for running in-database—transparent to the user. With AutoML, OML4Py supports automatic algorithm selection, feature selection, and model tuning to enhance user productivity as well as model accuracy and performance.

Oracle Machine Learning for R (OML4R) makes the open source R statistical programming language and environment more scalable through integration with Oracle Database. Data scientists and R users can take advantage of third-party packages from 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. OML4R extends R’s capabilities through three primary areas: (i) transparent exploration and preparation of database data from R through overloaded R function-to-SQL translation, (ii) R interfaces to in-database machine learning algorithms, and (iii) ease of deployment of user-defined R functions with embedded R execution through a SQL API.
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 widely used, free, integrated development environment that simplifies database development and management of Oracle Database in both on-premises and Oracle Database Cloud Service deployments. Oracle Data Miner users can share their analytical workflows and generate SQL scripts for 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 Spark (OML4Spark) provides machine learning in the Data Lake environment against big data clusters and is supported by Oracle R Advanced Analytics for Hadoop, a component of the Oracle Big Data Connectors. OML4Spark provides an R API to manipulate data stored in the local file system, HDFS, HIVE, Impala, Spark DataFrames, Oracle Database, and other JDBC sources. OML4Spark supports both native scalable algorithms as well as popular Apache Spark MLlib algorithms that are integrated in the same R framework.

By combining the functionality of multiple special-purpose databases—such as machine learning, in-memory, blockchain, JSON, text, graph, spatial, partitioning, sharding—with Oracle’s converged database strategy, users reduce complexity and management for an end-to-end optimized experience. Oracle Machine Learning benefits from this strategy with minimized data movement, cross-technology usage, reduced total cost of ownership, and enabling the fastest way to deliver enterprise-wide data-driven solutions.
Executive Summary

Oracle's mission is “to help people see data in new ways, discover insights, and unlock endless possibilities.” As such, Oracle provides enterprises with a broad range of capabilities within the Oracle Machine Learning product family, for data scientists, data and business analysts, data engineers and IT, application and dashboard developers, and executives to leverage their investments in Oracle technology and discover new insights and make predictions.

Large and small businesses alike know the value of applying machine learning technology to solve key business problems and achieve competitive, if not leadership position in their industries. Oracle enables democratizing machine learning across the enterprise and further the ability of businesses to leverage powerful machine learning tools.

Move the algorithms, not the data

With Oracle Machine Learning, Oracle “moves the algorithms to the data” by delivering machine learning algorithms that process the data within Oracle Database and on big data platforms, where the data reside, both on premises or in the Cloud. This approach minimizes or eliminates data movement, leverages the scalability of Oracle Database, preserves data security, and accelerates time from data acquisition to model deployment. Oracle delivers parallelized in-database and Spark-based implementations of machine learning algorithms and integration with the leading open source environments R and Python. Oracle Machine Learning delivers the performance, scalability, security, and automation required by enterprise-scale data science projects – both on premises and in the Cloud. In-database algorithms specifically leverage the parallelism, scalability, and security features of the industry’s leading database.

Oracle Database - a converged database

Users should not have to create and manage multiple databases to access different types of data and different analytical functionality, which adds complexity and cost. Instead, all such functionality should exist in a single converged, multi-model database, bringing together a broad set of algorithms that can operate on data with various data types and data models.

In the world of databases, the Oracle Converged Database can support all types of data and processing. Relational Data, Document Data, Spatial Data, Text Data and Graph Data can all be efficiently stored and processed in the same database. OLTP, Real-Time Analytics, Machine Learning, IoT, and Blockchain workloads can all efficiently run in the same database. There is no longer any need to deploy, manage, and synchronize many specialized databases to use these technologies.

Like Smartphones, the Oracle Converged Database not only supports these data types and workloads, it does it BETTER than the specialized databases. Further, the synergy of having all these capabilities together in one database enables new possibilities like real-time Machine Learning based fraud detection directly in the OLTP system. Also, using declarative open SQL rather than the bespoke languages needed by specialized databases makes it dramatically simpler for developers to implement these technologies.

–Juan Loaiza, Oracle

Empower enterprise users with SQL, R, and Python

Polls¹ show that in 2019, Python, R and SQL were among the most popular languages used for machine learning projects. Different business problems and their underlying data require different analytical techniques and algorithms to be successful. Similarly, different teams and team members have a variety of skills often focused on a particular language. Oracle Machine Learning supports all three popular data science / machine learning languages so users

can leverage the language that best suits them and the problem at hand, including access to over 30 in-database machine learning algorithms. Oracle Machine Learning supports a natural evolution from Oracle data professionals to data scientists within the same platform. In-database models are first class database objects with many of the same data management features available for other database objects like tables and views, such as permissions and auditing. This enables enterprises to deploy machine learning results for use in dashboards and applications, including Oracle tools such as Oracle Analytics Cloud and Oracle APEX, whether from SQL or REST interfaces.

By supporting R and Python, data scientists and other R and Python users access Oracle Database as a high-performance compute engine for scalable data exploration and preparation with native R and Python functions and syntax. The SQL and PL/SQL interface to in-database algorithms are exposed through well-integrated R and Python interfaces. Users can deploy user-defined functions to the database script repository, invoking R functions from SQL, and Python functions from a REST API. OML4R is available with for Oracle Database and Oracle Database Cloud Service, and OML4Py is currently available with Oracle Autonomous Database.

**Automation for increased productivity and model quality**

A significant part of the machine learning process involves iteration – performing feature engineering and selection, trying various algorithms and variations of their settings to find the best model, tuning the algorithm hyperparameters, and comparing and selecting the highest ranking model for a given business problem.

To increase data scientist productivity and enable non-expert data scientists to leverage machine learning, Oracle Machine Learning introduces automated machine learning, or AutoML in both the Python interface and a no-code user interface. AutoML can reduce the overall compute time required to derive a high-quality model. AutoML also makes powerful in-database algorithms more accessible to citizen data scientists as they do not need to know algorithm hyperparameter details, algorithm-specific data transformations, and write code to compare and select models according to various score metrics.

*Figure 1: Oracle Machine Learning AutoML UI aids data scientist productivity while enabling non-expert users to develop in-database machine learning models and deploy them immediately to OML Services.*
OML4Py AutoML supports the explicit construction of the modeling pipeline using three main capabilities: algorithm selection, feature selection, and model tuning. Oracle Machine Learning AutoML UI builds on OML4Py’s AutoML functionality to provide a no-code, easy-to-use interface that requires minimal specification, just the data table and target column to predict. Users benefit from built-in intelligence to produce and deploy better models more quickly and easily through a fully assembled AutoML experiment pipeline.

Oracle Machine Learning in-database algorithms also support automation at the algorithm level: automatic data preparation for algorithm-specific transformations of numeric and categorical data, integrated text mining for token and theme extraction, and automated partitioned models that produce a convenient-to-use ensemble model. Oracle Data Miner facilitates data assembly and transformation, and model building and evaluation of multiple algorithms (through Oracle Database 12c) in an easy-to-use drag-and-drop graphical workflow representation for on-premises and Oracle Database Cloud Service users.

**Production deployment - the critical step**

Some might refer to production deployment as the Achilles’ heel of data science projects. Consider that 87% of data science projects never make it to production ([VentureBeat.com](https://venturebeat.com)). Further, [Gartner](https://www.gartner.com) notes that “Through 2022, only 20% of analytic insights will deliver business outcomes.” When application developers or IT try to integrate machine learning models or open source R and Python scripts in production, they are faced with the realities of addressing the tedious and error-prone task of model transformation from one language to another, and handling backup, recovery, data security, and scalability concerns. By having the algorithms in Oracle Database, these steps and concerns are virtually eliminated. With Oracle Machine Learning, deployment is immediate – in-database models exist in the database and can be invoked from SQL queries, R and Python scripts that are stored in Oracle Database and run in automatically spawned and controlled R and Python engines. As such, Oracle greatly simplifies production deployment by providing the “plumbing” so enterprise teams can focus on their machine learning solutions.

OML Services enables application development by providing REST endpoints to deployed machine learning models. OML Services supports model deployment and model management via REST APIs for both in-database OML models and models produced by third-party machine learning algorithms. Oracle Machine Learning Services is available with Oracle Autonomous Database and can manage exported in-database models Oracle Database, Oracle Database Cloud Service, and Oracle Autonomous Database. These model management and deployment services extend to third-party models exported in Open Neural Networks Exchange (ONNX) format. Such third-party models may have been produced separately, such as in Oracle Cloud Infrastructure Data Science, from packages such as scikit-learn, Tensorflow, PyTorch, among others. OML Services also supports cognitive text analytics for extracting topics and keywords, and performing sentiment analysis, text summary and similarity.

As enterprises amass greater data volumes and supplement corporate data with external data sources, the ability to integrate and prepare data at scale for machine learning is essential. When it comes to deployment, the ability to streamline the time from business problem definition to solution deployment is critical to deliver ROI on machine learning and data science projects. Oracle Machine Learning customers have achieved impressive results, including:

- **UK National Health Service (NHS)**, the system of public healthcare providers in the United Kingdom, were able to assemble and analyze billions of data points on prescriptions, medicines, medical exemptions, doctor relationships, and call center services from across the organization and use this to reveal potential new efficiencies and effective treatments to provide better outcomes. NHS achieved very fast ROI – 581 million British pounds or $717 million USD in the first 2 years – and now over 1.5 billion British pounds or nearly $2 billion USD within 5 years and counting. Nina Monckton, head of Information Services, NHS Business Services Authority, states “We chose Oracle because the solution could cope with very large data volumes running into billions of rows and could scale as volumes increase. In addition, the Oracle solution required no IT team support to run the queries, which enables our team of data analysts to be self-sufficient. Oracle Exalytics’ in-memory capability gave us the speed we required, and Oracle’s engineered systems accelerated deployment and reduced risk. Working with Oracle has been a very positive experience. The team has been
incredibly responsive and provided a number of experts to help us get up and running as quickly as possible. With one vendor providing the whole solution, it’s very easy for us. If we need help, we know where to go.”

- CaixaBank, a major financial services company in Spain, integrates data from bank branches, ATMs, and internet and mobile banking to gain a complete understanding of customers and offer personalized banking solutions. CaixaBank gains customer loyalty and business competitiveness with improved messaging to reach customers more effectively, better informing them of new bank services and products to boost sales.

- Certegy, a leader in check payment services of more than $2B for over 4,000 clients and 23 of top 50 US retailers, chose Oracle Autonomous Data Warehouse to provide the repository for 850M records for statistical modeling, risk assessments and fraud detection by data scientists. Certegy uses Oracle Analytics, Oracle Machine Learning, and Oracle APEX for improved integration, data quality, and performance than Microsoft PowerBI and Python. Using Oracle Machine Learning, Certegy achieves more accurate, real-time risk scores with an expected 10% reduction in fraud and improved customer service.

- StubHub, the world’s largest ticket marketplace, uses Oracle Machine Learning in-database models and integrated R capabilities to run real-time fraud detection models in their database. With this solution, StubHub reduced online fraud by 90%. This not only saves money, but also significantly improves the customer experience. Of course, fraudsters notice when they are shut down and so change their way of operating. This requires reacting quickly, or fraud goes up. With Oracle Machine Learning, StubHub can react very quickly; updating their predictive model happens with the application and database still running. If administrators notice a problem in the morning, they are able to fix it in an hour, deploy the new model immediately, with no downtime.

For more details about these and other customer stories, see OML Customers.
What is Machine Learning?

Machine learning is the process of *automatically* processing large volumes of data to find hidden patterns, discover new insights, and make predictions for data-driven problems including:

- 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 referred to as *predictive analytics* or *data mining*, as a technology has been delivering measurable value for decades. Today, machine learning is even more pervasive—being implemented and deployed across enterprises globally. As big data analytics technologies and user adoption matures and expands, machine learning use cases and integrated “intelligent” applications that push “the art of the possible” emerge every day and constantly raise the bar for user’s expectations.

Move the Algorithms, Not the Data

Data is big; algorithms are small. Hence, it makes sense to move the algorithms to the data rather than move the data to the algorithms. Oracle realized this challenge of big data and the algorithms required to make sense of it when it acquired machine learning technology and the development team from Thinking Machines Corporation in 1999. Oracle capitalized on core Oracle Database strengths, specifically aggregation, parallelism, scalability, and the database architectural framework to build models and score data with all computations inside the database kernel. Neither the data, the machine learning models, nor the predictions and insights need to leave the database.

Today, Oracle Machine Learning supports over 30 in-database algorithms, integrated text mining, automated data preparation, and automated machine learning (AutoML), among other innovative features, and extends Oracle Database to enable users to build intelligent applications and analytics dashboards.

What is Oracle Machine Learning?

Oracle Machine Learning provides support for data driven problems by offering a wide range of powerful data exploration and preparation capabilities, as well as machine learning algorithms implemented inside Oracle Database and for Spark-based big data platforms. The in-database algorithms are exposed directly through SQL, as well as Python and R APIs. The Spark-based big data algorithms, discussed later, are native highly parallel and scalable, and exposed through an R API.

Oracle Machine Learning algorithms leverage the underlying SQL features and can even analyze data in its original star schema representation including standard structured tables and views, transactional data and aggregations, unstructured data as found in character large object (CLOB) data types—using Oracle Text to extract tokens. Through Oracle's converged database, users have a single environment for combining the machine learning with both spatial and graph analytics.

Oracle Machine Learning in-database algorithms take advantage of database parallelism for both model building and scoring, honor security and user privilege schemes, and adhere to audit tracking database features. In Exadata (on premises, Public Cloud, or Cloud at Customer) and Autonomous Database, OML in-database models are scored at the storage tier using Oracle’s *smart scan* technology.
Oracle Machine Learning supports both simple and complex machine learning 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.

**Support for SQL and open source R and Python**

SQL has been the standard language for data management for over 40 years. For data analytics and machine learning, however, various languages compete—R, Python, SAS, MATLAB, and others. These have been long time favorites, but in recent years, open source Python and R have surged to the top of the pack. As noted earlier, per the KD Nuggets data mining industry community annual polls, Python, R, and SQL are the top 3 languages and Oracle Machine Learning supports all three with OML4Py, OML4R, and OML4SQL.

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. Python and R are widely popular and powerful open source programming languages and statistical environments, are free, and taught in most data science educational programs. Data analysts, data scientists, researchers, and academics often start by learning to use Python or R. This leads to a pool of experts who can now work with their data inside Oracle Database—reaping the benefits of in-database performance while using Python and R open source languages.

**In-Database Processing with Oracle Machine Learning**

Oracle Machine Learning extends Oracle Database as a data science platform with powerful in-database analytics and algorithms are performed directly on database data. Results, insights, and real-time machine learning models are available and managed in Oracle Database.

An Oracle Machine Learning in-database model is a first-class database schema object, built by invoking a PL/SQL procedure or corresponding OML4R or OML4Py functions. This model can then be used to score data via built-in OML4SQL, OML4R, or OML4Py functions. When building models, Oracle Machine Learning leverages standard scalable database technology, including parallel execution, bitmap indexes, and aggregation techniques. Oracle Machine Learning also takes advantage of custom-built technologies, including parallel infrastructure, IEEE float, Intel

---

**Technical Brief / Oracle Machine Learning / Version 1.0**

Copyright © 2021, Oracle and/or its affiliates / Public
Math Kernel Library (MKL), automatic data preparation for binning, missing value handling, and support for unstructured text data.

The power of in-database algorithms as SQL functions is even more evident when scoring machine learning models, whether in batch or in online transactional processing (OLTP) environments. Scoring data with these models is amazingly fast—millions of records in seconds. Using the smart scan technology of Oracle Exadata Storage Cells, SQL predicates and OML models get pushed down to the storage layer for processing. In both cases, only those records that satisfy the predicates are pulled from disk for further processing inside the database.

For example, here is a simple SQL query to find the US customers likely to churn:

```sql
SELECT cust_id
FROM customers
WHERE region='US' AND prediction_probability(churn_model,'Y' using *) > 0.8;
```

Reading this query, it equates to customers with more than 80% probability to churn based on the model called `churn_model`. The scoring computations are pushed down to storage via predicate functions using Oracle's proprietary Exadata Smart Scan technology for 2-5 times faster scoring performance than non-Smart Scan in-database scoring.

Whether building models or scoring data, in-database machine learning enables users to streamline their data science projects from days and weeks to seconds, minutes, or hours.

**Automation drives productivity**

The data science process can be highly iterative. Popular methodologies, such as the CRoss Industry Standard Process for Data Mining (CRISP-DM), have iterative loops built in, especially for data preparation, and model building and evaluation. Fortunately, a significant part of this process can be automated. Such automation not only improves data scientist productivity by addressing some of the more time iterative aspects of machine learning, but also enables non-experts to leverage machine learning even if they do not know the finer details of the algorithms and their settings. Still, such automation does not eliminate the need for competent machine learning skills.

Automation provided in Oracle Machine Learning includes automatic data preparation (ADP), automated text processing, automated partitioned models, and in OML4Py and OML AutoML UI, automated machine learning (AutoML). AutoML provides algorithm and feature selection, and model selection and tuning, and in OML AutoML UI, adaptive sampling. These features all help reduce the effort, and in some cases, the knowledge required to go from business problem to business solution using Machine Learning.

**Integrating Open Source R and Python with Oracle Database**

Oracle Machine Learning for R (OML4R) and Oracle Machine Learning for Python (OML4Py) make the open source R and Python languages and environment more scalable through integration with Oracle Database. The strengths of these languages include drawing on a rich open source ecosystem, with an extensive array of powerful and extensible graphics and statistical packages. Typical R and Python users, however, face challenges including memory constraints, single threaded processing, and difficulty deploying or translating R and Python model logic into enterprise applications or dashboards.

OML4Py integrates Python with Oracle Database and overloads select Python functions to equivalent SQL and in-database algorithms—enabling Python users to operate on database-resident data without using SQL. The overloaded functions form the transparency layer, which transparently converts requested Python functionality into equivalent SQL for in-database processing. Through the transparency layer, Python programmers can create pandas DataFrame proxy objects that reference database tables and views, enabling them to access, analyze, and manipulate data that resides in the database. The database automatically optimizes the corresponding SQL code to improve query efficiency, and takes advantage of column indexes, table partitioning, and database parallelism.
With embedded Python execution, users can run user-defined Python functions in one or more system-spawned and controlled Python engines. User-defined functions are stored in the database script repository. Data scientists and analysts can develop, refine, and deploy Python code that leverages database parallelism and scalability as well as in-database algorithms without having to learn SQL. User-defined Python functions can be invoked in a non-parallel, data-parallel, or task-parallel manner. Similar capabilities are provided for R users with OML4R.

OML4R is supported through Oracle Database on premises and Oracle Database Cloud Service. With OML4R embedded R execution, users can extend Oracle Machine Learning’s native functionality by creating user-defined functions that use third-party open source R packages, such as those from CRAN, that are installed at the database server machine. Embedded R execution functions can be invoked using both R and SQL APIs. R users can leverage their favorite R IDE, such as RStudio, to connect to an Oracle Database instance and use OML4R.

OML4Py is supported on Oracle Autonomous Database and accessed through Oracle Machine Learning Notebooks. With OML4Py embedded Python execution, users can define user-defined functions that include native Python functionality and select pre-installed packages. Embedded Python execution functions can be invoked using both Python and REST APIs.
Figure 4: Oracle Machine Learning Notebooks supports Python and SQL interpreters. In this notebook, the Python user leverages the Expectation Maximization clustering algorithm, displaying a scatterplot of cluster assignments using the matplotlib library.

**Oracle Machine Learning AutoML User Interface**

Oracle Machine Learning AutoML User Interface provides a no-code user interface supporting automatic machine learning. OML AutoML UI supports both data scientist and citizen data scientist productivity, as well as non-expert user access to powerful in-database algorithms for classification and regression.

OML AutoML UI automates model building with minimal user input – just specify the data and the target in what is called an “experiment” and the interface does the rest. For those users wanting more control, there are advanced settings to limit the set of algorithms tried and the number of models ranked.

As shown below, running an experiment produces a leaderboard of models and their performance, and allows users to select models for deployment in OML Services, or use models directly through SQL and Python. Users can also generate a notebook containing the OML4Py code that generates that model.

**Oracle Machine Learning Services**

Oracle Machine Learning Services extends OML functionality to support model deployment and model management via a REST API for both in-database OML models and models exported in Open Neural Networks Exchange (ONNX) from third-party algorithms. OML Services provides REST endpoints available with Oracle Autonomous Database. The REST endpoints enable storing machine learning models along with their metadata and creating scoring endpoints for each model.

In-database models supporting classification, regression, clustering, and feature extraction can be deployed to OML Services. Third-party models supporting classification and regression, including those from packages such as scikit-learn, TensorFlow, PyTorch, among others, can be deployed to OML Services provided they can be exported in ONNX format.

In addition, OML Services supports proprietary cognitive text capabilities, including topic discovery, keywords extraction, summarization, sentiment analysis, and similarity. Initial languages supported include English, Spanish,
and French (based on a pre-built Wikipedia-based Explicit Semantic Analysis model), with sentiment analysis for English in the first release.

Through the ONNX format third-party model deployment feature, users can import cognitive image third-party models for image classification using images or tensors.

The set of REST API capabilities in OML Services is shown in Figure 5.

![Figure 2: REST API capabilities provided by Oracle Machine Learning Services](image)

**Oracle Data Miner—a SQL Developer Extension**

Oracle Data Miner, an extension to the desktop application Oracle SQL Developer, is designed for data scientists and citizen data scientists who may prefer a drag-and-drop user interface to work directly with data inside the database using a workflow paradigm. Oracle Data Miner workflows capture and document the user’s machine learning methodology. Users can save and share their workflows with others to automate and publish their machine learning methodologies. Oracle Data Miner is available for Oracle Database on premises and Oracle Database Cloud Service, with support for a select set of in-database algorithms.

Users quickly learn Oracle Data Miner to visualize, explore, prepare and transform data, build and evaluate machine learning models, and use extensive model and model evaluation viewers. Then, they can apply OML in-database models to new data or generate SQL and PL/SQL scripts to deploy analytical workflow.

These scripts can be passed to database developers for deployment within the same or a different Oracle Database. Application developers can programatically run workflows using the Oracle Data Miner PL/SQL workflow API or just run the appropriate SQL and PL/SQL scripts. This enables an easy integration of predictive methodologies into applications for wider use throughout the enterprise.
Oracle Machine Learning in Data Lakes

“Boil down” the data

Big data usually represent vast amounts of structure and unstructured data that is collected more easily than it is analyzed, and it is often stored in Hadoop clusters, on so-called data reservoirs or data lakes. However, for the purposes of solving data-driven problems, oftentimes the main interest is not in the raw data itself, but on a summary of the data provided by counts, percentages, and “events” to be joined with other curated data inside Oracle Database. The resulting data can usually be used to solve a business problem using machine learning. The idea is to “boil down” the data lake into new engineered features that capture domain expert knowledge for producing better models.

The data environment outside Oracle Database introduces new data management and data analysis challenges. Oracle Cloud SQL addresses this challenge by extending SQL processing to Hadoop via Oracle Big Data Service, and to Oracle Object Storage via Autonomous Database. Using smart scan technology developed for Exadata, Cloud SQL pushes down SQL logic to operate on Hive external tables or Cloud Object Storage data 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, boil them down, and combine them with data of known value managed inside a database or data warehouse.

Data stored in Hadoop and Object Storage is often voluminous and has a sparse representation, such as transactional format. Given that much of the data may come from sensors, Internet of Things, “tweets”, and other high-volume sources, Cloud SQL is a powerful complementary tool to Oracle Machine Learning for aggregating data at various levels, e.g., counts, maximum values, minimum values, thresholds counts, averages, and 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 perform machine learning in-database using Oracle Machine Learning.

**Oracle Machine Learning for Spark**

Another option is to perform machine learning directly against the big data cluster using Oracle Machine Learning for Spark (OML4Spark), a component of the Big Data Connectors. The OML4Spark R API provides functions for manipulating data stored in a local file system (e.g., as CSV files), 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 linear model, generalized linear model (GLM), and multi-layer perceptron (MLP) Feed Forward Neural Network algorithms that run atop Spark infrastructure. OML4Spark native algorithms scale better and run faster than the open-source alternatives of Apache Spark MLib. OML4Spark wraps Spark MLib algorithms in R.
functions within the OML4Spark framework using the R formula specification and Distributed Model Matrix data structure for greater ease of use. The Spark MLib R functions in OML4Spark can be run either on a Hadoop cluster using YARN (to dynamically form a Spark cluster), a dedicated standalone Spark cluster, or even on a local Spark session.

Cloud SQL and OML4Spark can be combined from Oracle Database or Oracle 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.

**Oracle Machine Learning—the Fastest Way to Deliver AI Applications**

Oracle’s strategy of making big data and big data analytics simple makes it easier to develop, refine and deploy machine learning-enhanced, intelligent applications. Data, user access, security and encryption, scalability, applications development environment, and machine learning are processed in Oracle Database and the Oracle ecosystem. Now, it is easier than ever to add predictions and real-time actionable insights to your enterprise applications, dashboards, or tools that can speak SQL, REST, or other APIs to Oracle Database.

Oracle has been developing intelligent 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 Configure, Price, Quote
- Oracle Depot Repair
- Oracle Spend Classification
- Oracle Adaptive Intelligence Foundation for Anti-Money Laundering
- Oracle Adaptive Access Management’s Identity Management
- Oracle Retail Science Platform Cloud Service
- Oracle Retail XBRI Loss Prevention
- Oracle Retail Customer Segmentation
- Oracle Cloud Infrastructure Operations Insights
- Oracle Capacity Planning

![Figure 5: Oracle HCM Predictive Workforce application delivers pre-built Oracle Machine Learning predictive analytics for employee attrition, employee performance and “What if?” analysis.](image)

**Conclusion**
Oracle Machine Learning delivers scalable, parallelized, machine learning algorithms within Oracle Database and Big Data environments as part of Oracle’s converged database strategy. Oracle Machine Learning provides machine learning algorithms supporting classification, regression, clustering, association rules, feature extraction, time series, ranking, row importance, and anomaly detection, and that are available through multiple interfaces, including OML Notebooks, OML AutoML UI, popular language APIs for SQL, R and Python, OML Services, Oracle Data Miner, and inside big data environments on Apache Spark via OML4Spark.

Oracle Machine Learning algorithms that run in Oracle Database take full advantage of Oracle Database scalability, parallelism, security, and structured and unstructured data processing capabilities. This makes Oracle the powerful platform for developing machine learning solutions and embedding them in enterprise applications either on premises or on Oracle Cloud. Oracle’s multiple decades of leading-edge data management experience combines with the strategy of “moving algorithms, not data.” By integrating data management and machine learning into the converged Oracle Database, Oracle reduces total cost of ownership, eliminates data movement, and enables fast deployment to enterprise dashboards and applications while enabling data scientists, citizen data scientists, data analysts, and application developers.

For more information

Website

Blogs: Machine Learning and R Technologies

Documentation

Office Hours Sessions

Oracle Machine Learning LiveLabs
Connect with us

Call +1.800.ORACLE1 or visit oracle.com. Outside North America, find your local office at: oracle.com/contact.

blogs.oracle.com    facebook.com/oracle    twitter.com/oracle

Copyright © 2021, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

This device has not been authorized as required by the rules of the Federal Communications Commission. This device is not, and may not be, offered for sale or lease, or sold or leased, until authorization is obtained.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0120

Disclaimer: If you are unsure whether your data sheet needs a disclaimer, read the revenue recognition policy. If you have further questions about your content and the disclaimer requirements, e-mail REVREC_US@oracle.com.