ORACLE DATABASE 12c

Oracle's Machine Learning and Advanced Analytics Data Management Platforms

Move the Algorithms; Not the Data

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" are driving companies to change. Just to keep pace, they must learn new skills and implement new practices that leverage those new data sources and technologies. Increasing customer expectations from sharing their digital exhaust with corporations in exchange for improved customer interactions and greater perceived value are pushing companies forward. Big data and analytics offer the promise to satisfy these new requirements. Cloud, competition, big data analytics and next-generation "predictive" applications are driving companies towards achieving new goals of delivering improved "*actionable insights"* and better outcomes. Traditional BI & Analytics approaches don't deliver these detailed predictive insights and simply can't satisfy the emerging customer expectations in this new world order created by big data and the cloud.

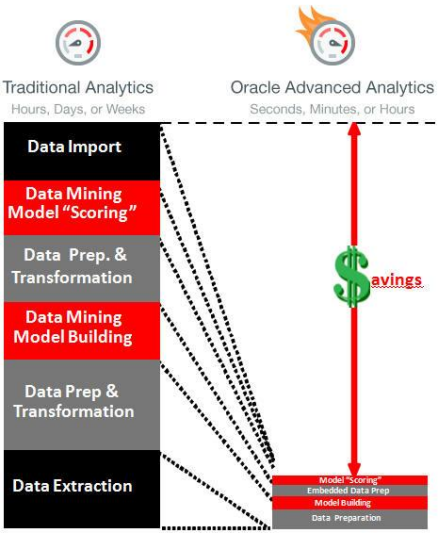
Unfortunately, with big data, as the data grows and expands in the three V's; velocity, volume and variety (data types), new problems emerge. Data volumes grow and data becomes unmanageable and immovable. Scalability, security, and information latency become new issues. Dealing with unstructured data, sensor data and spatial data all introduce new data type complexities.

Traditional advanced analytics has several information technology inherent weak points: data extracts and data movement, data duplication resulting in no single-source of truth, data security exposures, separate and many times, depending on the skills of the data analysts/scientists involved, multiple analytical tools (commercial and open source) and languages (SAS, R, SQL, Python, SPSS, etc.). Problems become particularly egregious during a deployment phase when the worlds of data analysis and information management collide.

Traditional data analysis typically starts 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 especially designed for statisticians and data scientists to analyze data. The analytics they perform range from simple descriptive statistical analysis to advanced, predictive and prescriptive analytics. If

a data scientist builds a predictive model that is determined to be useful and valuable, then IT needs to be involved to figure out deployment and enterprise deployment and application integration issues become the next big challenge. The predictive model(s)—and all its associated data preparation and transformation steps—have to be somehow translated to SQL and recreated inside the database in order to apply the models and make predictions on the larger datasets maintained inside the data warehouse. This model translation phase introduces tedious, time consuming and expensive manual coding steps from the original statistical language (SAS, R, and Python) into SQL. DBAs and IT must somehow "*productionize*" these separate statistical models inside the database and/or data warehouse for distribution throughout the enterprise. Some vendors will charge for specialized products and options for just for predictive model deployment. This is where many advanced analytics projects fail. 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, analyze, export, move load paradigm. With Oracle Database 12c and the Oracle Advanced Analytics Option, big data management and big data analytics are designed into the data management platform from the beginning. Oracle's multiple decades of R&D investment in developing the industry's leading data management platform, Oracle SQL, Big Data SQL, Oracle Exadata, Oracle Big Data Appliance and integration with open source R are seamlessly combined and integrated into a single platform—the Oracle Database.



Oracle Advanced Analytics eliminates data movement and combines big data management with big data analytics.

Oracle's vision is a big data and analytic platform for the era of big data and cloud to:

- **Make big data and analytics simple** *(for any data size, on any computer infrastructure and any variety of data, in any combination)* ***and***
- **Make big data and analytics deployment simple** *(as a service, as a platform, as an application)*

Oracle Advanced Analytics offers a wide library of powerful in-database algorithms and integration with open source R that together can solve a wide variety of business problems and can be accessed via SQL, R or GUI. Oracle Advanced Analytics, an option to the Oracle Database Enterprise Edition 12c, 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, and text mining and sentiment analysis. Oracle Advanced Analytics empowers data analyst, data scientists and business analysts to more extract knowledge, discover new insights and make informed predictions—working directly with large data volumes in the Oracle Database.

Data analysts/scientists have choice and flexibility in how they interact with Oracle Advanced Analytics. Oracle Data Miner is an Oracle SQL Developer extension designed for data analysts that provides an easy to use “drag and drop” workflow GUI to the Oracle Advanced Analytics SQL data mining functions (Oracle Data Mining). Oracle SQL Developer is a free integrated development environment that simplifies the development and management of Oracle Database in both traditional and Cloud deployments. When Oracle Data Miner users are satisfied with their analytical methodologies, they can share their workflows with other analysts and/or generate SQL scripts to hand to their DBAs to accelerate model deployment. Oracle Data Miner also provides a PL/SQL API for workflow scheduling and automation.

R programmers and data scientists can use the familiar open source R statistical programming language console, RStudio or any IDE to work directly with data inside the database and leverage Oracle Advanced Analytics’ R integration with the database (Oracle R Enterprise). Oracle Advanced Analytics’ Oracle R Enterprise provides transparent SQL to R translation to equivalent SQL and Oracle Data Mining functions for in-database performance, parallelism, and scalability—this making R ready for the enterprise.

Application developers, using the ODM SQL data mining functions and ORE R integration can build completely automated predictive analytic solutions that leverage the strengths of the database and the flexibly of R to integrate Oracle Advanced Analytics analytical solutions into BI dashboards and enterprise applications.

By integrating big data management and big data analytics into the same powerful Oracle Database 12.2c data management platform, Oracle eliminates data movement, reduces total cost of ownership and delivers the fastest way to deliver enterprise-wide predictive analytics solutions and applications.

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. Companies have been data mining large volumes of data for years. What's been new and more challenging is the increasing pace of the "big data" volumes, velocities and varieties of sources coupled with new customer expectations of what new "actionable insights" can be achieved. This 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.

Unfortunately, as big data grows and expands over time in its three V's; velocity, volume and variety, new problems emerge. Data volumes grow and eventually become near immovable. Eventually at some point, it becomes impractical to move large data amounts to separate servers for the data analysis. During the big data explosion, many problems are experienced such as data movement, data duplication, security, creation of "data analysis sprawl-marts", separation of data management from data analysis and worse, information latency expands, oftentimes to multiple days and weeks.

Traditional data analysis methods contribute to these problems. Data analysts and data scientist typically have their own special "tools" that they've learned to use (SAS, R, SPSS or Python, etc.) so require data extracts from the database /data warehouse, transforms and loads to dedicated, separate analytical servers. If a data scientist builds a "good" predictive model, then a new problem emerges. Deployment of that model to where and when it is most needed and integration into applications e.g. BI dashboards, call centers, websites, ATMs and mobile devices becomes the next big challenge for IT. The predictive model(s)—and all the associated data preparation and transformation steps—have to be recreated in the destination platform to make the predictions on the larger data tables. For Oracle environments, this export, data analysis, import results outer loop complicates the data analysis unnecessarily and introduces the time consuming and expensive model deployment phase. IT is asked to "*productionize*" the models and re-implement them using SQL inside the database.

The challenge is that the models were originally created using a statistical programming language (SAS, R, SPSS and Python.) but to productionize them, they must run as SQL functions inside the database. This is where the big time sink occurs and errors can be introduced. For organizations who strive to be leaders, efficient data collection, data management, analysis, and deployment of predictive models, insights and actionable business intelligence are the keys to their success. Traditional data analysis methods just won't suffice. Add Hadoop, sensor data, tweets, and ever expanding new data reservoirs and the whole problem just gets worse.

Predictive Analytics!

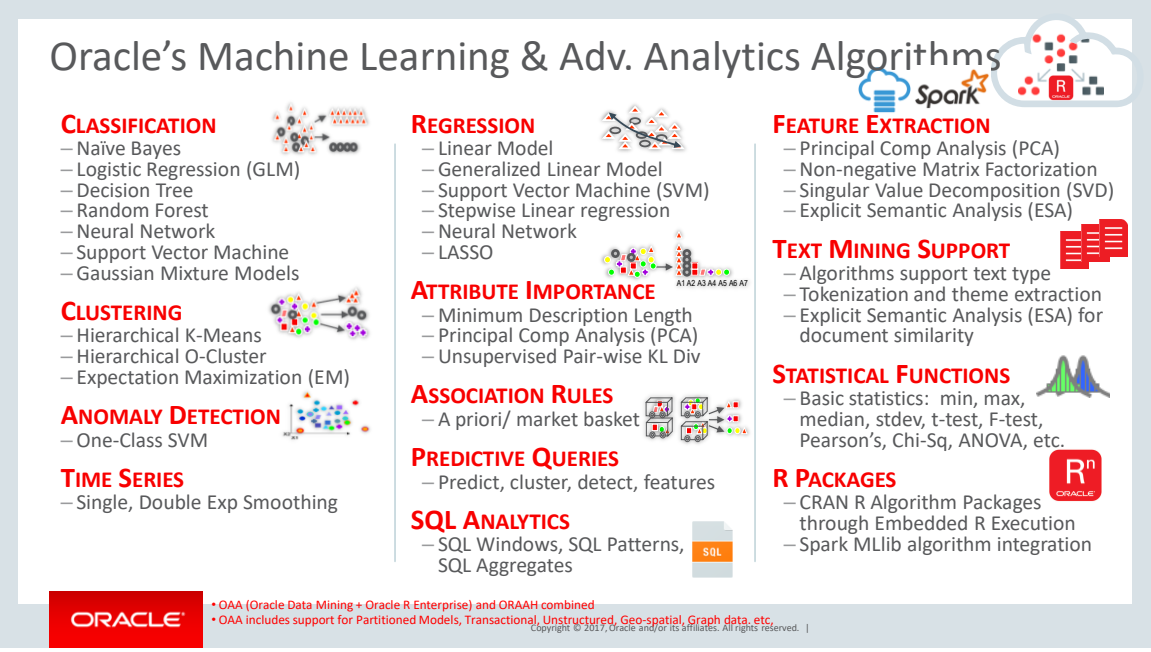
Predictive analytics is the process of automatically sifting through large amounts of data to find previously hidden patterns, discover valuable new insights and make informed predictions for data-driven problems such as:

- 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

Predictive Analytics as a technology has been delivering measurable value for years. Predictive Analytics climbed it's was up *Gartner's Hype Cycle for Emerging Technologies* and reached the Gartner's enviable "plateau of productivity" in 2013. Today in 2015, predictive analytics are being implemented and deployed in enterprises and applications ranging from predicting churn and employee turnover, to flagging medical fraud and tax non-compliance to targeted selling and real-time recommendation engines. As big data analytics technologies and user adoptions evolve, mature and expand, predictive analytics use cases and integrated "predictive" applications that push "the art of the possible" are emerging every day and are constantly raising the bar of new user expectations.

Oracle Advanced Analytics provides support for these data driven problems by offering a wide range of powerful workhorse data mining algorithms that have been implemented in a relational database environment (RDBMS). Algorithms are implemented as SQL functions inside the database. Oracle Advanced Analytics' data mining algorithms hence 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 i.e. CLOB data types (using Oracle Text to parse out "tokens") and spatial data. Oracle Advanced Analytics in-database SQL data mining functions take advantage of parallelism inside the database for both model build and model apply, honor all security and user privilege schemes, adhere to revision control and audit tracking database features and can mine data in its native and potentially encrypted form inside the Oracle Database.



Oracle Advanced Analytics 12.2c data mining functions are implemented as SQL functions that can be accessed by SQL, PL/SQL, R and the Oracle Data Miner GUI.

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 *moving 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 cutting edge machine learning algorithms and statistical functions as native SQL functions with full SQL language support. With Oracle Advanced Analytics, data mining 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*i*, Oracle's first data mining algorithms took advantage of available core Database's strengths—specifically, counting, parallelism, scalability and other database architectural underpinnings. Essentially, the first two Oracle data mining algorithms, Naïve Bayes and A Priori algorithms, 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 leave the database.

OAA Naïve Bayes algorithm can quickly builds 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?" Let's take an example in a bit more detail for comprehension. Let's say we are interested in selling Product A (e.g. a motorcycle or $500 shoes, etc.). The Oracle Advanced Analytics data mining algorithms, specifically the Naïve Bayes algorithm, of all the customers who purchased Product A, it counts how many customers were male vs. female. How many rent an apartment vs. owns their own home? How many have children and how many? Each of these answers 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.

OAA's A Priori "market basket analysis" algorithm counts items in each customer's transactional "baskets" while looking for co-occurring items e.g. A + B appear together frequently, and then provides conditional probability AR rules. For example:

> **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 Advanced Analytics, a store could decide to place the milk near the cereal and bananas, 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 types of ways that big data analytics can find "actionable insights" from data. Obviously, more data, more advanced analytics methodologies and fast enterprise wide deployment can open new doors to many new big data and analytics applications and solutions possibilities.

SQL and R Support

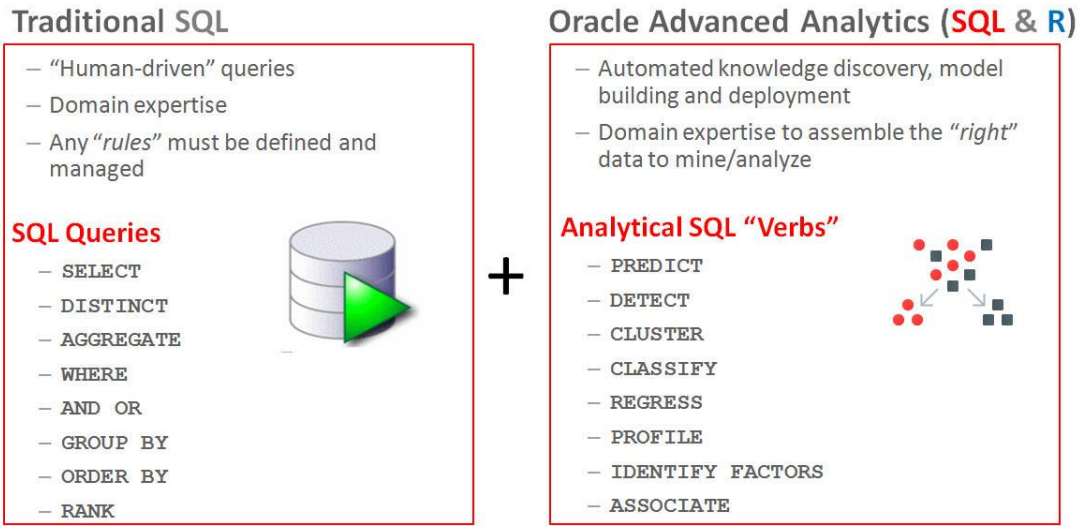
Where SQL is the standard language for data management and has been for 40+ years, for data analysis, various languages compete—R, SAS, Python and SQL and others. SAS, S+, SQL, SPSS and Matlab have been long time favorites, but in recent past years, open source R especially has surged to the top of the pack and Python and others have emerged. Per the KDD 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 Advanced Analytics supports both languages—SQL and R. There are legions of developers who know SQL for data management and Oracle provides support for data mining and advanced analytics via Oracle Advanced Analytics' SQL data mining functions and provides tight, industry leading integration with open source R statistical programming language.

Most Oracle customers are very familiar with SQL as a language for query, reporting, and analysis of structured data. It is the de facto standard for analysis and the technology that underlies most BI tools. R is a widely popular open source programming language for statistical analysis that is free and because of that is 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 the Oracle Database using either SQL or R languages.

Over the past decade and one-half, Oracle Advanced Analytics has matured and has been developed to now in Oracle 12c, the Oracle Advanced Analytics Option delivers nearly twenty scalable, parallelized, in-database implementations of workhorse predictive analytics algorithms. Oracle Advanced Analytics exposes these data mining algorithms as SQL functions that are accessible via SQL, R language and the Oracle Data Miner GUI, an extension to Oracle SQL Developer for the most common data driven problems e.g. clustering, regression, prediction, associations, text mining, associations analysis, etc. All Oracle Advanced Analytics algorithms are implemented deep inside the database and take full advantage of the Oracle Database's industry leading scalability, security, SQL functions, integration, ETL, Cloud, structured, unstructured and spatial data types features and strengths and can be accessed via both SQL and R—and GUI.

Hence, you can think of Oracle Advanced Analytics like this…

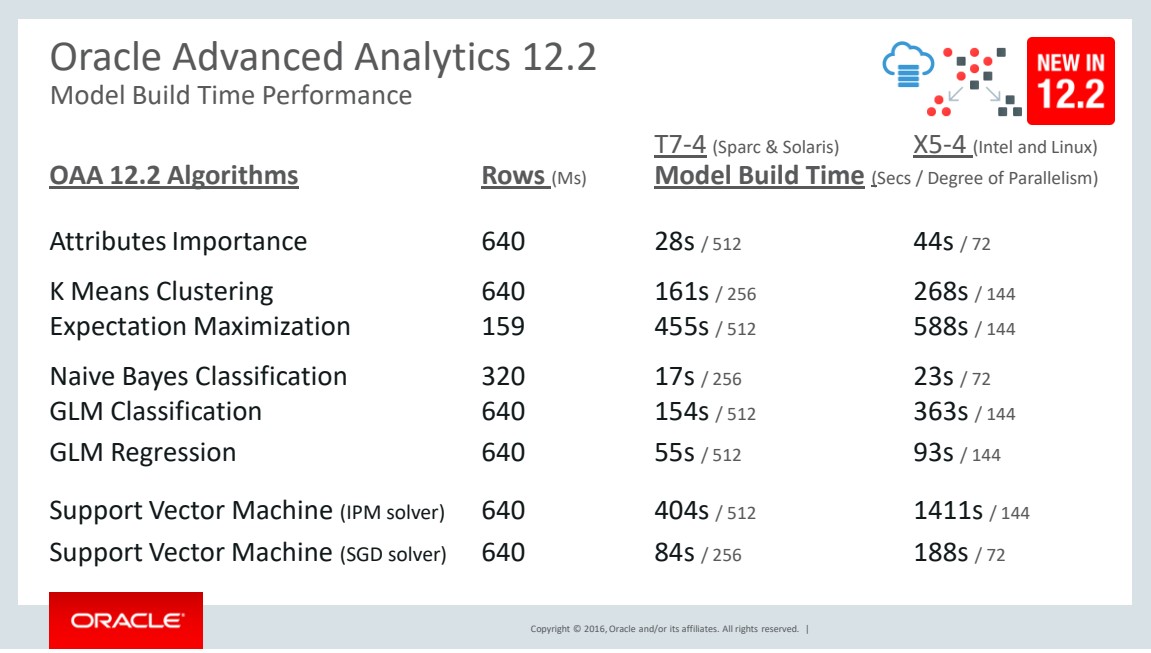


Oracle Advanced Analytics extends the SQL language to add powerful analytical verbs e.g. predict, detect, associate, cluster.

In-Database Processing with Oracle Advanced Analytics

Oracle Advanced Analytics extends the database into a comprehensive advanced analytics platform for big data analytics. 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.

A data mining model is a schema object in the database, built via a PL/SQL API that prepares the data, learns the hidden patterns to build an OAA model which can then be scored via built-in OAA data mining SQL functions. When building models, Oracle Advanced Analytics leverages existing scalable technology (e.g., parallel execution, bitmap indexes, aggregation techniques) and additional developed new Oracle Advanced Analytics and Oracle Database 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, etc.).



Oracle Advanced Analytics 12.2 delivers significant performance improvements and can build machine learning models on hundreds of millions of records and hundreds of attributes in seconds or minutes.

The true power of embedding data mining functions within the database as SQL functions is most evident when scoring data mining models. Once the models have been built by learning the hidden patterns in the historical data, applying the models to new data inside the database is blazingly fast. Scoring is then just a row-wise function. Hence, Oracle Advanced Analytics can "score" many millions of records in seconds and is designed to support online transactional processing (OLTP) environments.

Using Exadata's "smart scan" technology it gets better. With Oracle Advanced Analytics running in an Exadata environment, SQL predicates and OAA predictive models get pushed down to the hardware layer for execution.

- For Oracle Exadata environments, pushed to Exadata storage level for execution
- For Oracle Big Data Appliance (Hadoop) environments, pushed to BDA storage level 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:

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

Scoring function executed in Exadata or on BDA

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

Typically, in order to perform proper analysis on data, analysts have to make explicit decisions about how to "bin" data, deal with missing values and oftentimes reduce the number of variables (feature selection) to be used in the models. Over the past 15 years, Oracle Advanced Analytics has evolved and now can automate most of the steps typically required in data mining projects. Today, Automated Data Preparation (ADP) automatically bins numeric attributes using default and user customizable binning strategies e.g. equal width, equal count, user-defined and similarly bins categorical attributes into N top values and "other" or user-defined bins. Missing values are automatically replaced by a statistical value (i.e. mean, median, mode, etc.) instead of that record being removed from the analysis. ADP is used both for model building and then again for applying the models to new data. Users can of course override ADP settings if they choose.

Oracle Advanced Analytics provides support for attribute reduction (Attribute Importance using the Minimum Description Length algorithm) and feature reduction techniques (Principal Components Analysis and Non-Negative Matrix Factorization). However, *each* of the Oracle Advanced Analytics algorithms (e.g. Decision Trees, Generalized Linear Regression, Support Vector Machines, Naïve Bayes, K-Means Clustering, Expectation Maximization Clustering, Anomaly Detection 1-Class SVMs, etc.) has their own built-in automated strategies for attribute reduction and selection so the an explicit variable reduction step is optional, but not necessary. Users of course can control algorithm and data preparation settings or accept the intelligent defaults.

Transactional data, e.g. purchases, transactions, events, etc. represent much of the data that is important to build good predictive models. Oracle Advanced Analytics mines this data in its native transactional form and leverages the database's aggregation functions to summarize it and then feed vector of the data (e.g. item purchases) and join it to other customer 2-D data to provide a 360 degree customer view. Oracle Advanced Analytics models, e.g. classification, regression and clustering models, ingest this aggregated transactional attribute as a "nested table". Deep inside the Oracle Advanced Analytics' in-database processing, records are processed as triplets: Unique_ID, Attribute_name, and Attribute_value. That's just part of the secret sauce of how Oracle Advanced Analytics leverages the core strengths of the Oracle Database. Market basket analysis would of course mine this data in its native transactional data form (typically not aggregated) to find co-occurring items in baskets.

Unstructured data i.e. text is also processed in a similar fashion inside the database. Oracle Advanced Analytics uses Oracle Text's text processing capabilities and multi-language support to "tokenize" any CLOB data type e.g. text, Word, Adobe Acrobat, etc. As Oracle Text is a free feature in every Oracle Database, Oracle Advanced Analytics leverages it to pre-process unstructured data to then feed vectors of words and word coefficients (TFIDF—term frequency inverse document frequency) into the algorithms. Oracle Advanced Analytics just treats the unstructured attributes as additional input attributes e.g. police comments, physician's notes, resume, emails, article, abstract, etc. that get joined with everything else (e.g. Age, Income, Occupation, etc.) that is being fed into the Oracle Advanced Analytics data mining algorithms. Spatial data, web clicks and other data types can also be joined and included in Oracle Advanced Analytics data mining models.

Oracle Data Miner Workflow GUI; a SQL Developer extension

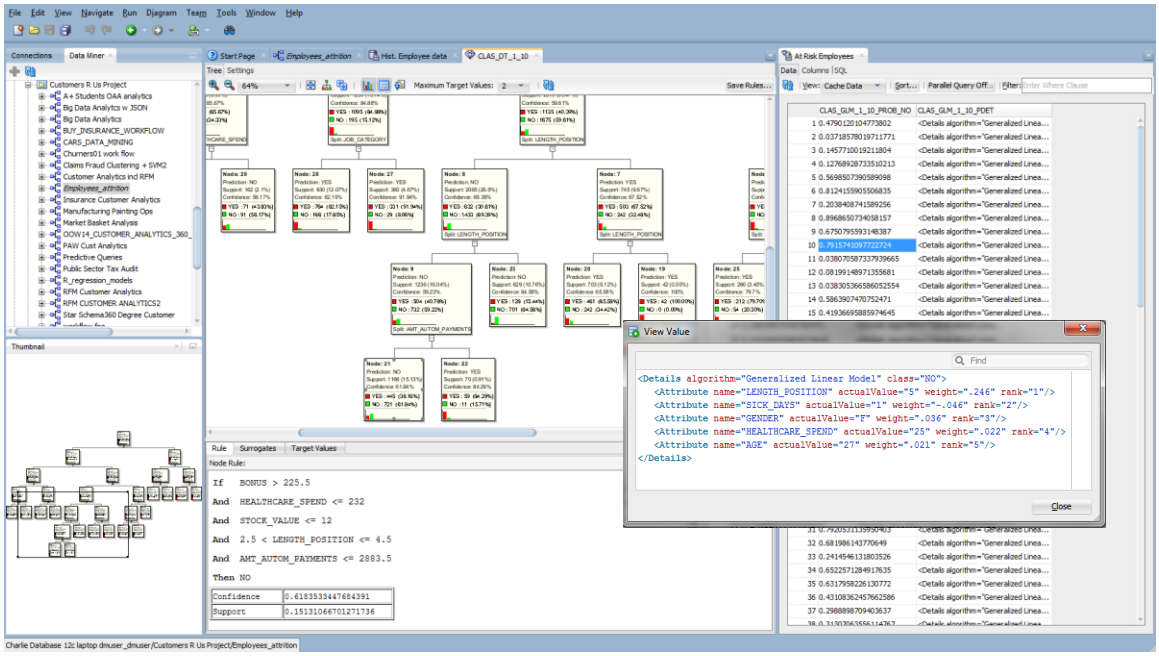
Oracle Data Miner GUI, an extension to Oracle SQL Developer 4.2, is designed for users who prefer an easy to use GUI for their data analysis and don't necessarily want know have to know how to program in either SQL or R—or just don't want to write code. Oracle Data Miner enables data analysts, business analysts and data scientists to work directly with data inside the database using Oracle Data Miner's graphical "drag and drop" workflow paradigm.

Data analysts easily learn how to use Oracle Data Miner and can quickly visualize and explore the data graphically, prepare and transform their data as necessary, build and evaluate multiple data mining models using extensive

model viewing and model evaluation viewers. Then they can apply Oracle Data Mining models to new data for deployment and/or they can generate SQL and PL/SQL scripts to deploy Oracle Data Mining's predictive models throughout the enterprise.

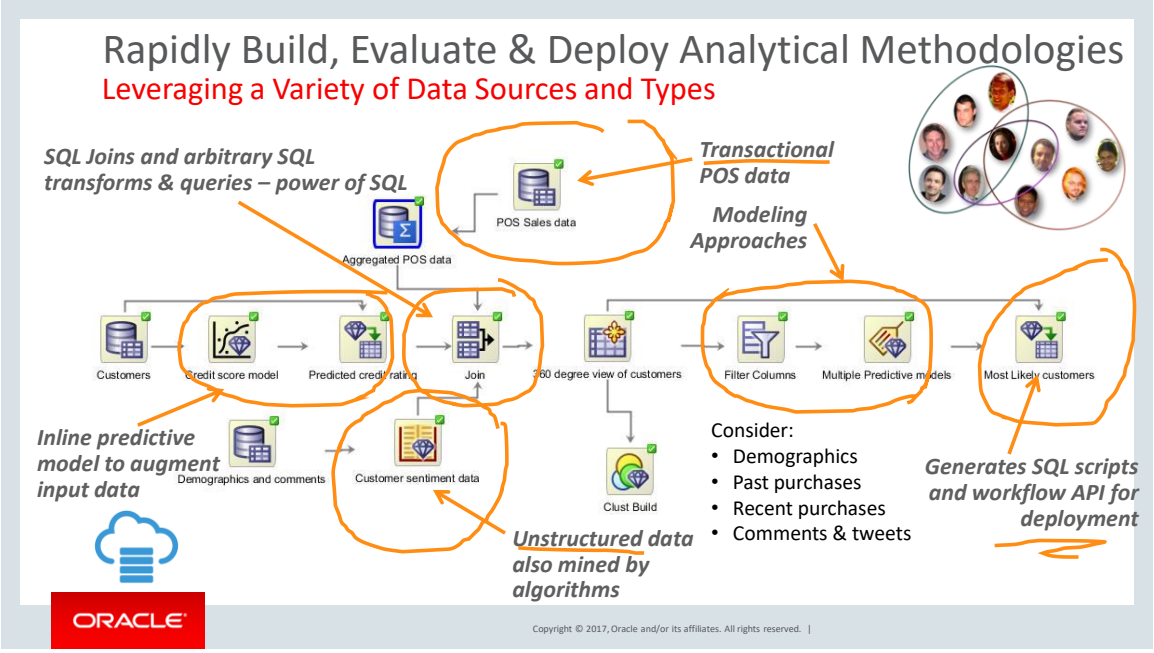
Oracle Data Miner work flows capture and document the user's analytical methodology and can be saved and shared with others to automate and publish advanced analytical methodologies.

When the data analysts are done, Oracle Data Miner generates SQL scripts to pass to their DBAs for immediate deployment using the Oracle Database for combined data management and predictive analytics. Application developers can programmatically call the workflows using the Oracle Data Miner PL/SQL workflow API to fully automate the discovery and distribution of new business intelligence and actionable insights and to integrate 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 for data analysts to explore their data, build, evaluate and apply predictive models and deploy advanced analytical methodologies as SQL and PL/SQL scripts.

Data analysts can use Oracle Data Miner to experiment and assemble very simple to complex advanced analytical methodologies. For example, a data analyst 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 different classification, regression or anomaly detection models for better accuracy and usefulness.



Data analysts can quickly build simple to sophisticated analytical methodologies that mine data they have access to in the Oracle Database. All data, models and results remain inside the database.

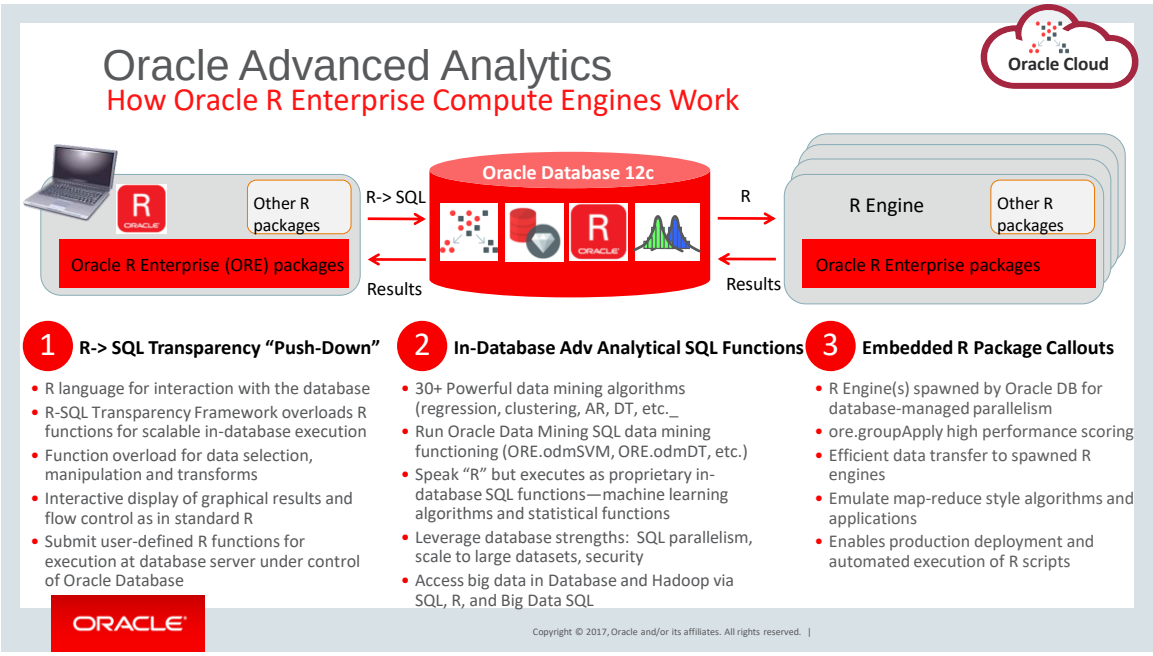
Oracle R Enterprise—Integrating Open Source R with the Oracle Database

Oracle R Enterprise, a component of the Oracle Advanced Analytics Option, 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 are that it is *free*—open source, powerful and extensible, has an extensive array of graphical and statistical packages and is constantly being expanded by the R user community who author and contribute R "packages". R's challenges are that it is memory constrained, single threaded, runs an outer loop that can slow down processing and is not generally considered to be "industrial strength". Contributed R packages are of varying quality.

Oracle R Enterprise integrates R with Oracle Database and maps R functions to equivalent SQL and Oracle Data Mining SQL functions and is designed for problems involving large amounts of data. It is a set of R packages (ORE) and Oracle Database features that enable an R user to operate on database-resident data without using SQL and to execute R scripts in one or more embedded R engines that run on the database server. Data analysts and data scientists can develop, refine, and deploy R scripts that leverage the parallelism and scalability of the database and the SQL data mining functions to automate data analysis in one step—without having to learn SQL

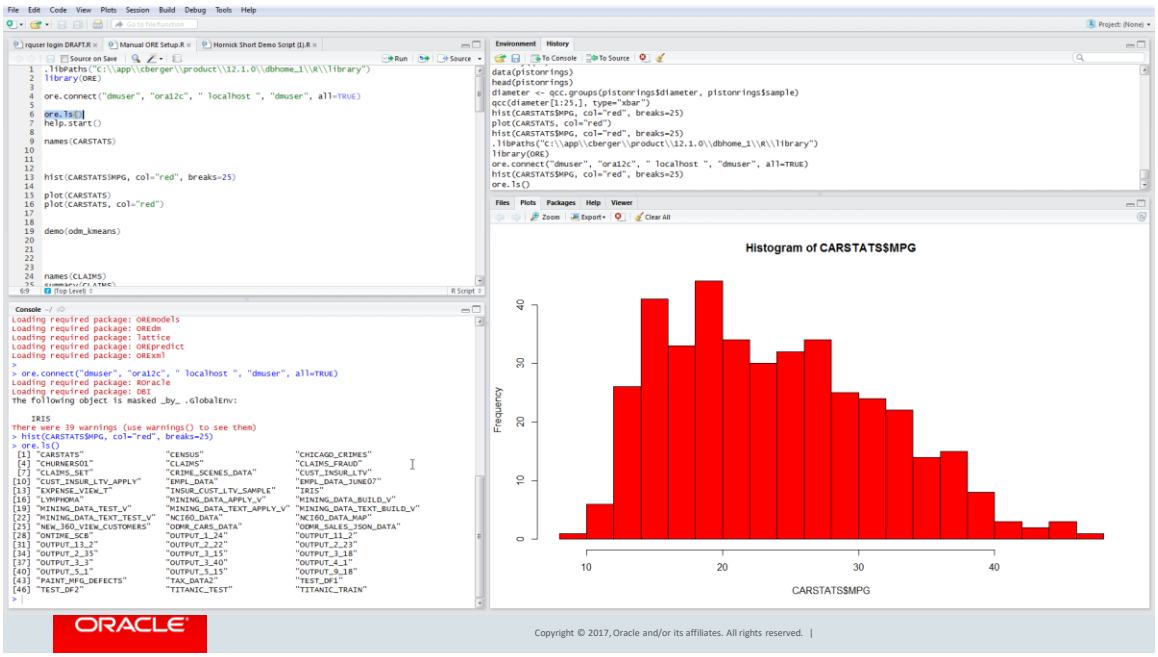
Oracle R Enterprise has overloaded open source R methods and functions that transparently convert standard R syntax into SQL. These methods and functions are in ORE packages that implement the Oracle R Enterprise transparency layer. With these functions and methods, R programmers can create R objects that access, analyze, and manipulate data that resides in the database. The database automatically optimizes the SQL code to improve the efficiency of the query. Oracle R Enterprise performs function pushdown for in-database execution of base R, Oracle SQL statistical functions, Oracle Data Mining SQL functions and selected popular R packages. Because it runs as an embedded component of Oracle Database, Oracle R Enterprise can run any R package either by function pushdown or via "embedded R mode" while the database manages the data served to the R engines. This

"embedded R mode" ability allows developers to extend Oracle Advanced Analytics' natively supported toolkit with any open source R packages and develop wide ranging and automated advanced analytics methodologies that are completely managed by the database.



Oracle Advanced Analytics' Oracle R Enterprise (ORE) component transparently pushes down R functions to equivalent in-database SQL functions for scalability and parallelism. ORE users can also leverage any R packages via "embedded R mode".

Users, who prefer to work in R to access and analyze their data, may use RStudio, or any R GUI, to connect to an Oracle Database and access Oracle Advanced Analytic's R integration (Oracle R Enterprise). Once a connection is made, the OAA/ORE session synchs the user's metadata so they see all their tables and views inside the database. When they run any base R language function it gets transparently mapped to equivalent SQL functions. R users using the OAA/ODM algorithms and OAA/ORE algorithms can perform scalable data mining in the database.

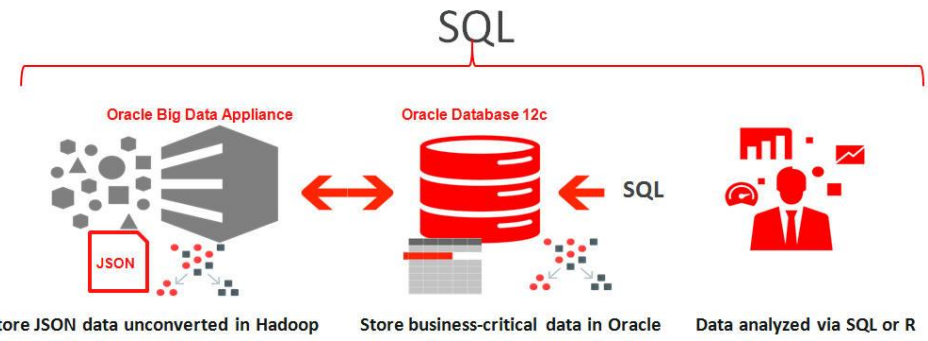


Oracle Advanced Analytics' Oracle R Enterprise component invoking in-database ODM algorithms (e.g., Support Vector Machine) from an RStudio console.

Hadoop, Oracle Big Data Appliance and Big Data SQL

Big data is now often stored in Hadoop servers. The separate data environment outside the database introduces new data management and data analysis challenges. Big Data SQL addresses this challenge by extending SQL processing to Hadoop via the Oracle Big Data Appliance. Using "smart scan" technology developed for Exadata, Big Data SQL pushes down SQL logic to operate on Hive tables. Data analysts can now more easily take advantage of new big data sources of *data of possibly unknown value* stored in big data reservoirs and combine that data with *data of known value* managed inside a database and/or data warehouse.

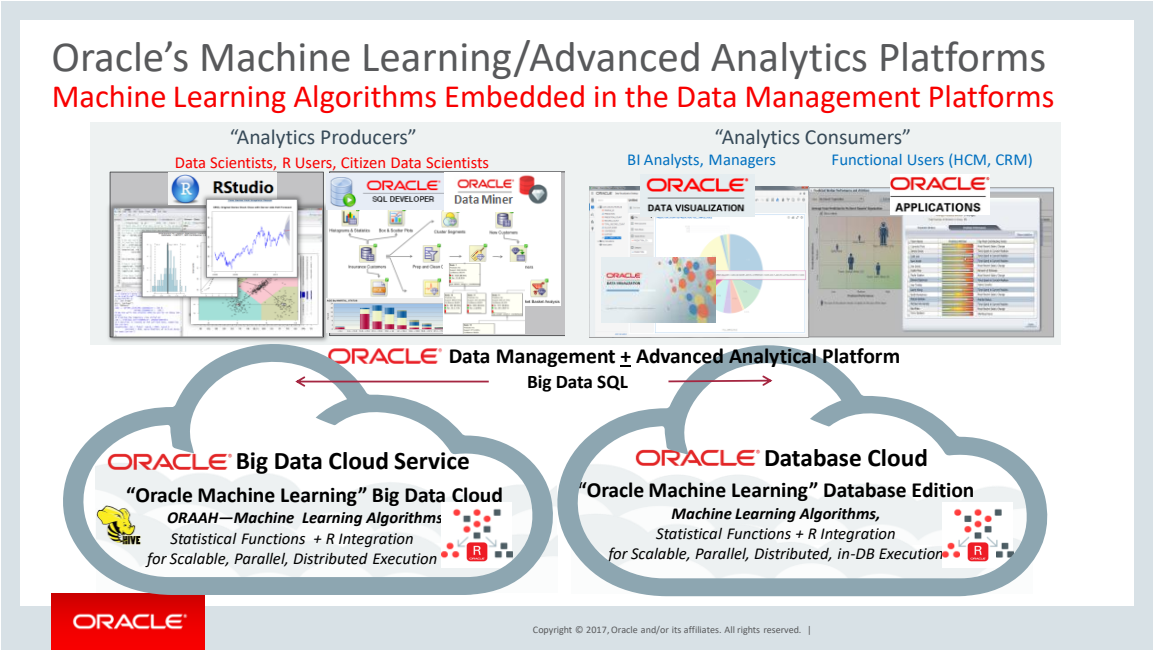
However, the data stored in Hadoop may be voluminous and sparse representation (transactional format) and lacking in information density. Given that much of the data may come from sensors, Internet of Things, "tweets" and other high volume sources, users can leverage Big Data SQL to collect counts, maximum values, minimum values, thresholds counts above or below user defined values, averages, shorter term averages and counts and longer time averages and counts, sliding SQL window averages and counts and comparisons of each to the other. So, filter "big data", reduce it, join it to other database data using Oracle Big Data SQL and then mine *everything* inside the Oracle Database using Oracle Advanced Analytics Option.



SQL and Big Data 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 Advanced Analytics.

A Platform for Developing Enterprise-wide Predictive Analytics Applications

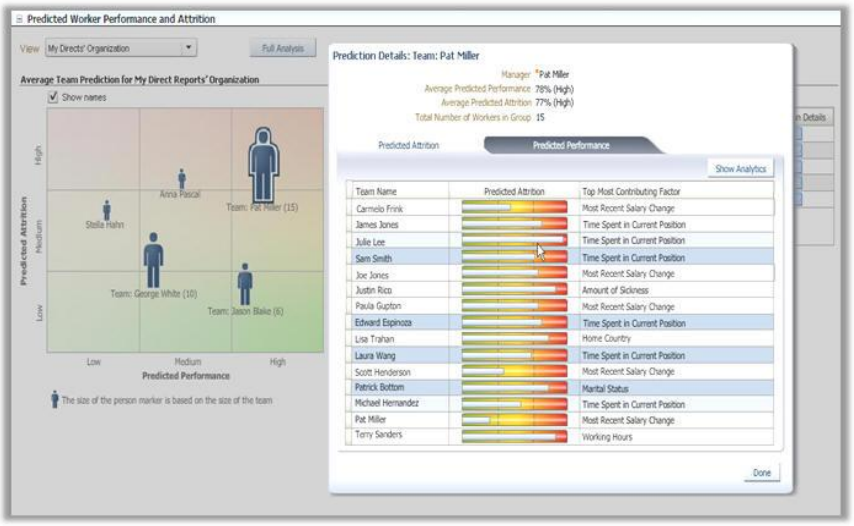
Oracle's strategy of making big data and big data analytics simple makes it easier to develop, refine and deploy predictive analytics applications—all is part of the database's functions. All the data, user access, security and encryption, scalability, applications development environment and powerful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



Oracle Advanced Analytics, a separately licensed feature of the Oracle Database, extends the database into a powerful analytical platform for both big data and big data analytics which is ideal for developing and deploying predictive analytics applications.

Oracle has been developing predictive analytics applications and provides next-generation predictive applications on premise and in the cloud including:

- Oracle Human Capital Management Predictive Workforce
- 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 HCM Predictive Workforce application delivers pre-built OAA predictive analytics for employee attrition, employee performance and *"What if?"* analysis.

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

Conclusion

Traditional BI and analytic approaches simply can't keep pace with requirements era of "big data" and "cloud". For organizations who strive to be leaders in their areas leveraging these new technologies, the prompt capture and collection of data of known and unknown value, the proper data management, assembly of relevant data and facile deep analysis and automation and deployment of the actionable insights is the key to success.

Oracle Advanced Analytics, a priced option to the Oracle Database 12.2c, collapses the traditional extract, move, load, analyze, export, move, load/import paradigm all too common today. Oracle Advanced Analytics delivers scalable, parallelized, in-database implementations of a wide library of workhorse predictive analytics algorithms (e.g. clustering, regression, prediction, associations, text mining, associations analysis, anomaly detection, etc.) as SQL functions within the Oracle Database 12c. Oracle Advanced Analytics exposes these predictive algorithms as SQL functions accessible via SQL (Oracle Data Mining OAA SQL API component), the Oracle Data Miner "drag and drop" workflow GUI, an extension to Oracle SQL Developer 4.2 and through tight integration w/ open source R (Oracle R Enterprise R integration component).

Because Oracle Advanced Analytics' in-database data mining machine learning/predictive analytics algorithms are built from the inside out of the Oracle Database and take full advantage of the Oracle Database's scalability, security, integration, cloud, structured and unstructured data mining capabilities, it make Oracle the ideal platform for big data + analytics solutions and applications either on-premise or on the Oracle Cloud.

With Oracle, data management and descriptive, predictive and prescriptive big data analytics are designed into the platform from the beginning. All of Oracle's multiple decades of leading edge data management and SQL and Big Data SQL is harnesses and combined with Oracle's design and development approach of *"moving the algorithms to the data"* vs. *"moving the data to the algorithms"*. Oracle's vision is to create a big data and analytic platform for the era of big data and the cloud to:

Make big data + analytics *simple:*

- Any data size, on any computer infrastructure
- Any variety of data, in any combination

Make big data and analytics deployment s*imple*

- As a service, as a platform, as an application

By integrating both big data management and big data analytics into a single unified Oracle Database 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 Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

ORACLE

CONNECT WITH US


blogs.oracle.com/oracle


facebook.com/oracle


twitter.com/oracle


oracle.com

Hardware and Software, Engineered to Work Together

White Paper Title
July 2017
Author: Charlie Berger, Sr. Director Product Management, Oracle Machine Learning & Advanced Analytics (charlie.berger@oracle.com)
Contributing Authors: [OPTIONAL]


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