JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Big Data Analytics with Oracle Advanced Analytics

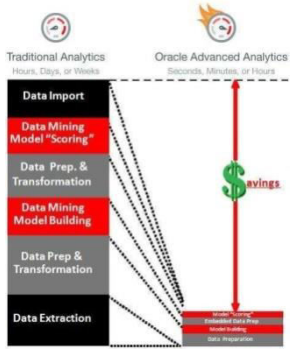
Dr. N. Thulasi, Professor,Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering and Technology

M.Parasakthi, Assistant Professor, Department of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology

ABSTRACT;

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.

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 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

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

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.

and valuable, then IT needsto

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, innovativeformsofprocessingforenhanceddecisionmaking,businessinsightsorprocessoptimization."*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 businesstheysupport e.g.marketing,customerservice,support,R&Dandoperations.Unfortunately,asbigdatagrowsandexpandsovertimeinitsthreeV's;velocity,volumeandvariety,newprobl ems 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 andweeks.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, datamanagement,analysis,anddeploymentofpredictivemodels,insightsandactionablebusinessintelligenceare the keys to their success. Traditional data analysis methods just won't suffice. Add Hadoop, sensor data, tweets, and ever expanding new data reservoirs andthe whole problem just gets worse

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.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Move the Algorithms, Not the Data;

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.Armedwiththesetypesofnewcustomerinsightsfrom OracleAdvancedAnalytics,astorecoulddecidetoplacethe milknearthecerealandbananas,offernewpromotional"breakfastkit"productbundlesormakereal-timecustomer 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 solutionspossibilities.

SQL and R Support;

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 analyticsviaOracleAdvancedAnalytics'SQLdataminingfunctionsandprovidestight,industryleadingintegration with open source R statistical programminglanguage. 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 implementeddeepinsidethedatabaseandtakefulladvantageoftheOracleDatabase'sindustryleadingscalability, security, SQL functions, integration, ETL, Cloud, structured, unstructured and spatial data types features and strengths and can be accessed via both SQL and R—andGUI.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

In-Database Processing with Oracle Advanced Analytics;

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.).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 applyingthemodelstonewdatainsidethedatabaseisblazinglyfast.Scoringisthenjustarow-wisefunction. Hence,OracleAdvancedAnalyticscan"score"manymillionsofrecordsinsecondsandisdesignedtosupport online transactional processing (OLTP)environments.

Automatic Data Preparation, Data Types, Star Schemas and "NestedTables";

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 DescriptionLengthalgorithm)andfeaturereductiontechniques(PrincipalComponentsAnalysisandNon-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 coursecancontrol algorithmanddatapreparationsettingsoracceptttheintelligentdefaults.

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 thedatabase'saggregationfunctionstosummarizeitandthenfeedvectorofthedata(e.g.itempurchases)andjoin it to other customer 2-D data to provide a 360 degree customer view. Oracle Advanced Analytics models, e.g. classification,regressionandclusteringmodels,ingestthisaggregatedtransactionalattributeasa"nestedtable".

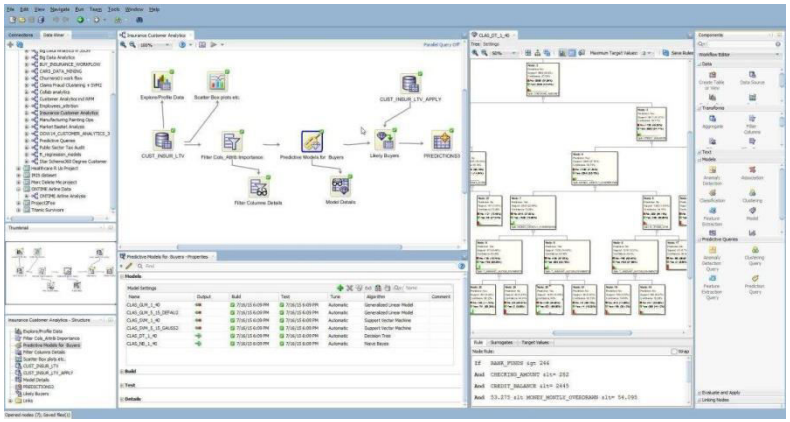
ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

DeepinsidetheOracleAdvancedAnalytics'in-databaseprocessing,recordsareprocessedastriplets:Unique_ID, Attribute_name, and Attribute_value. That's just part of the secret sauce of how Oracle Advanced Analytics leveragesthecorestrengthsoftheOracleDatabase.Marketbasketanalysiswouldofcourseminethisdatainits nativetransactionaldataform(typicallynotaggregated)tofindco-occurringitemsinbaskets.

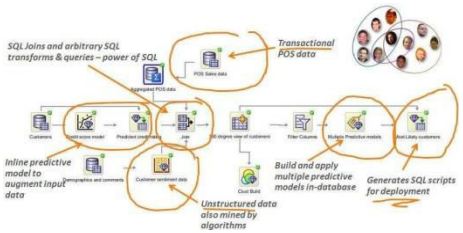
Oracle Data Miner Workflow GUI; a SQL Developer extension;

OracleDataMinerGUI,anextensiontoOracleSQLDeveloper4.1,isdesignedforuserswhopreferaneasytouse 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 workdirectlywithdatainsidethedatabaseusingOracleDataMiner'sgraphical"draganddrop"workflowparadigm

DataanalystseasilylearnhowtouseOracleDataMinerandcanquicklyvisualizeandexplorethedatagraphically, 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 theenterprise.

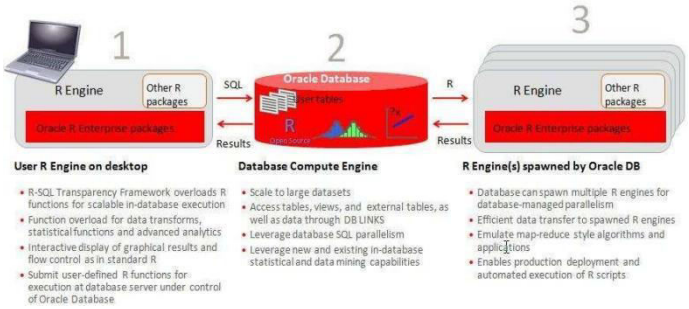


ISSN- 2394-5125 VOL 7, ISSUE 19, 2020



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.

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

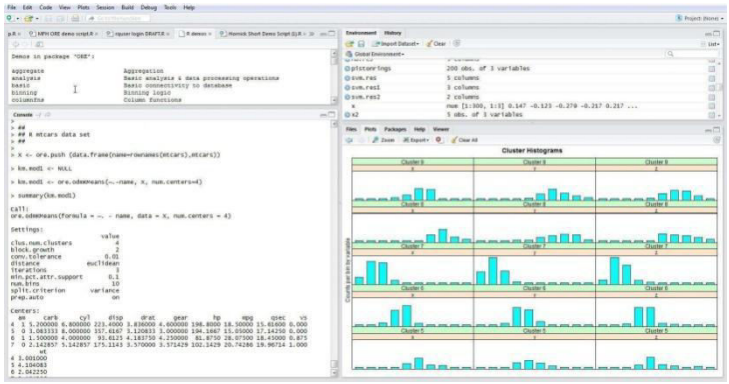


Oracle R Enterprise, a component of the Oracle Advanced Analytics Option, makes the open source R statistical programming language and environment ready for the enterprise and big data. “R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering,) andgraphicaltechniques,andishighlyextensible”(seehttps://www.r-project.org/).R’sstrengthsarethatitis*free*— opensource,powerfulandextensible,hasanextensivearrayofgraphicalandstatisticalpackagesandisconstantly 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 varyingquality.

Oracle R Enterprise integrates R with Oracle Database and maps R functions to equivalent SQL and Oracle Data MiningSQLfunctionsandisdesignedforproblemsinvolvinglargeamountsofdata.ItisasetofRpackages(ORE)andOracleDatabasefeaturesthatenableanRusertoope rateondatabase-residentdatawithoutusingSQLandto execute R scripts in one or more embedded R engines that run on the database server. Data analysts and

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

data scientistscandevelop,refine,anddeployRscriptsthatleveragetheparallelismandscalabilityofthedatabaseand theSQLdataminingfunctionstoautomatedataanalysisinonestep—withouthavingtolearnSQL

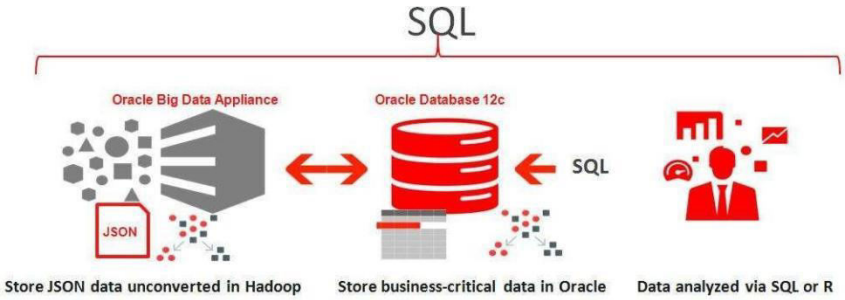


Hadoop, Oracle Big Data Appliance and Big Data SQL;

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 advantageofnewbigdatasourcesof*dataofpossiblyunknownvalue*storedinbigdatareservoirsandcombinethat datawith*dataofknownvalue*managedinsideadatabase and/or datawarehouse.

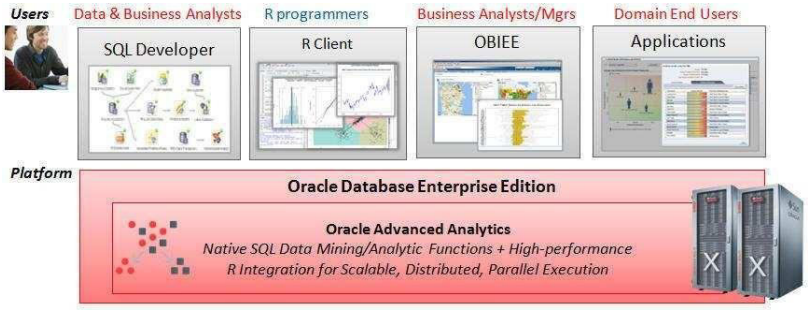
However, the data stored in Hadoop may be voluminous and sparse representation (transactional format) and lacking in information density. Given that much of the data may come from sensors, Internet of Things, “tweets” and other high volume sources, users can leverage Big Data SQL to collect counts, maximum 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.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020



A Platform for Developing Enterprise-wide Predictive AnalyticsApplications;

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.



ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Conclusion

TraditionalBIandanalyticapproachessimplycan'tkeeppacewithrequirementseraof"bigdata"and"cloud".For organizations who strive to be leaders in their areas leveraging these new technologies, the prompt capture and collectionofdataofknownandunknownvalue,theproperdatamanagement,assemblyofrelevantdataandfacile deepanalysisandautomationanddeployment oftheactionableinsightsisthekeytosuccess.

Oracle Advanced Analytics, a priced option to the Oracle Database 12c, 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 SQLfunctionsaccessibleviaSQL(OracleDataMiningOAASQLAPIcomponent),theOracleDataMiner"dragand drop" workflow GUI, an extension to Oracle SQL Developer 4.1 and through tight integration w/ open source R (Oracle R Enterprise R integrationcomponent).