An Oracle White Paper
March 2013

Big Data Analytics

Advanced Analytics in Oracle Database

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.

Executive Summary

Whether its fine-tuning supply chains, monitoring shop floor operations, gauging consumer sentiment, or any number of other large-scale analytic challenges, big data is having a tremendous impact on the enterprise. The amount of business data that is generated has risen steadily every year and more and more types of information are being stored in digital formats.

One of the challenges entails learning how to deal with all of these new data types and determining which information can potentially provide value to your business. It is not just access to new data sources, selected events or transactions or blog posts, but the patterns and inter-relationships among these elements that are of interest. Collecting lots of diverse types of data very quickly does not create value. You need analytics to uncover insights that will help your business. That's what this paper is about.

Big data doesn't only bring new data types and storage mechanisms, but new types of analysis as well. In the following pages we discuss the various ways to analyze big data to find patterns and relationships, make informed predictions, deliver actionable intelligence, and gain business insight from this steady influx of information.

Big data analysis is a continuum, not an isolated set of activities. Thus you need a cohesive set of solutions for big data analysis, from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems for ongoing analysis. Many organizations accomplish these tasks by coordinating the use of both commercial and open source components. Having an integrated architecture for big data analysis makes it easier to perform various types of activities and to move data among these components.

The Dawn of Big Data

Data becomes big data when its volume, velocity, or variety exceeds the abilities of your IT systems to ingest, store, analyze, and process it. Many organizations have the equipment and expertise to handle large quantities of structured data—but with the increasing volume and faster flows of data, they lack the ability to "mine" it and derive actionable intelligence in a timely way. Not only is the volume of this data growing too fast for traditional analytics, but the speed with which it arrives and the variety of data types necessitates new types of data processing and analytic solutions.

However, big data doesn't always fit into neat tables of columns and rows. There are many new data types, both structured and unstructured, that can be processed to yield insight into a business or condition. For example, data from twitter feeds, call detail reports, network data, video cameras, and equipment sensors often isn't stored in a data warehouse until you have pre-processed it to distill and summarize and perhaps to detect basic trends and associations. It is more cost effective to load the results into a warehouse for additional analysis. The idea is to "reduce" the data to the point that it can be put in a structured form. Then it can be meaningfully compared to the rest of your data, and scrutinized with traditional business intelligence (BI) tools.

Merging Traditional and Big Data Analysis

Taking advantage of big data often involves a progression of cultural and technical changes throughout your business, from exploring new business opportunities to expanding your sphere of inquiry to exploiting new insights as you merge traditional and big data analytics.

The journey often begins with traditional enterprise data and tools, which yield insights about everything from sales forecasts to inventory levels. The data typically resides in a data warehouse and is analyzed with SQL-based business intelligence (BI) tools. Much of the data in the warehouse comes from business transactions originally captured in an OLTP database. While reports and dashboards account for the majority of BI use, more and more organizations are performing "what-if" analysis on multi-dimensional databases, especially within the context of financial planning and forecasting. These planning and forecasting applications can benefit from big data but organizations need advanced analytics to make this goal a reality.

For more advanced data analysis such as statistical analysis, data mining, predictive analytics, and text mining, companies have traditionally moved the data to dedicated servers for analysis. Exporting the data out of the data warehouse, creating copies of it in external analytical servers, and deriving insights and predictions is time consuming. It also requires duplicate data storage environments and specialized data analysis skills. Once you've successfully built a predictive model, using that model with production data involves either complex rewriting of the model or the additional movement of large volumes of data from a data warehouse to an external data analysis server. At that point the data is "scored" and then the results are moved back to the data warehouse. This cycle of moving and re-purposing data to create actionable information can take days, weeks or even moths to complete.

While many organizations have achieved proficiency in exploiting their data through data analysis, they are still at the early stages of creating an analytic model that can deliver real business value from big data. The main obstacles are these slow and arcane processes for enabling direct and timely access to corporate data. However, new technologies are collapsing the old walls between IT and data analysts by enabling advanced analytics within the database itself, alleviating the need to move large volumes of data around.

At the same time, new types of data are supplementing traditional data sources and familiar BI activities. For example, weblog files track the movement of visitors to a website, revealing who clicked where and when. This data can reveal how people interact with your site. Social media helps you understanding what people are thinking or how they feel about something. It can be derived from web pages, social media sites, tweets, blog entries, email exchanges, search indexes, click streams, equipment sensors, and all types of multimedia files including audio, video, and photographic.

This data can be collected not only from computers, but also from billions of mobile phones, tens of billions of social media posts, and an ever-expanding array of networked sensors from cars, utility meters, shipping containers, shop floor equipment, point of sale terminals and many other sources.

Most of this data is less dense and more information poor, and doesn't fit immediately into your data warehouse. As we will see, some of it is better placed in Hadoop Distributed File System (HDFS) or in non-relational databases, commonly called NoSQL databases. In many cases, this is the starting point for big data analysis.

Techniques for Analyzing Big Data – A New Approach

When you use SQL queries to look up financial numbers or OLAP tools to generate sales forecasts, you generally know what kind of data you have and what it can tell you. Revenue, geography and time all relate to each other in predictable ways. You don't necessarily know what the answers are but you do know how the various elements

of the data set relate to each other. BI users often run standard reports from structured databases that have been carefully modeled to leverage these relationships.

Big data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a data model to define what each element means in the context of the others. There are several new issues you should consider as you embark on this new type of analysis:

- Discovery – In many cases you don't really know what you have and how different data sets relate to each other. You must figure it out through a process of exploration and discovery.
- Iteration – Because the actual relationships are not always known in advance, uncovering insight is often an iterative process as you find the answers that you seek. The nature of iteration is that it sometimes leads you down a path that turns out to be a dead end. That's okay – experimentation is part of the process. Many analysts and industry experts suggest that you start with small, well-defined projects, learn from each iteration, and gradually move on to the next idea or field of inquiry.
- Flexible Capacity – Because of the iterative nature of big data analysis, be prepared to spend more time and utilize more resources to solve problems.
- Mining and Predicting – Big data analysis is not black and white. You don't always know how the various data elements relate to each other. As you mine the data to discover patterns and relationships, predictive analytics can yield the insights that you seek.
- Decision Management – Consider the transaction volume and velocity. If you are using big data analytics to drive many operational decisions (such as personalizing a web site or prompting call center agents about the habits and activities of consumers) then you need to consider how to automate and optimize the implementation of all those actions.

For example you may have no idea whether or not social data sheds light on sales trends. The challenge comes with figuring out which data elements relate to which other data elements, and in what capacity. The process of discovery not only involves exploring the data to understand how you can use it but also determining how it relates to your traditional enterprise data.

New types of inquiry entail not only what happened, but why. For example, a key metric for many companies is customer churn. It's fairly easy to quantify churn. But why does it happen? Studying call data records, customer support inquiries, social media commentary, and other customer feedback can all help explain why customers defect. Similar approaches can be used with other types of data and in other situations. Why did sales fall in a given store? Why do certain patients survive longer than others? The trick is to find the right data, discover the hidden relationships, and analyze it correctly.

Big Data Use Cases

This section includes a few use cases that demonstrate the potential of big data analytics within various business domains.

Example #1: Machine-Generated Data

As the "Internet of Things" grows steadily each year, researchers predict that the amount of data generated by machines will one day outstrip the amount of data generated by humans. Machina Research, a UK-based research firm, believes there will be 12.5 billion "smart" connected devices—excluding phones, PCs and tablets—in the world in 2020, up from 1.3 billion today. Equipment sensors are prevalent in heavy machinery, automobiles, assembly lines, electric grids, computer equipment, and many other domains. And that's just the

beginning, as more and more devices are manufactured with sensors that monitor their own operation and log the results for troubleshooting and analysis. For example, manufacturing companies commonly embed sensors in their machinery to monitor usage patterns, predict maintenance problems, and enhance build quality. Even consumer devices such as bicycles, washing machines, and thermostats are part of this machine-to-machine (M2M) communications phenomenon.

Studying these data streams allows them to improve their products and devise more accurate service cycles. Electronic sensors not only monitor mechanical and atmospheric conditions, but also the biometrics of the human body. In health care there is a huge opportunity not only to improve patient outcomes but also to monitor trends in health care diagnoses, treatments, and claims to make better clinical and administrative decisions. The opportunities become even more compelling once data is analyzed in aggregate form. If a thousand sensors reveal a pattern of equipment failure, or a thousand cardiac monitors show a correlation between biometric levels and adverse reactions, then we can begin to turn trends into predictions – and ultimately use big data to take corrective or preemptive action.

Once again, finding the patterns is the key. For example, insurance companies are now asking drivers to voluntarily contribute data that tracks their movement, locations, and where they are at various times of the day so they can develop better risk profiles for each customer. By showing that they drive the speed limit, travel in areas that incur fewer accidents, and avoid high crime areas customer can qualify for a lower cost insurance plan.

Example #2: Online Reservations

If you were running an online travel booking website, there are lots of interesting things you could do with your data to better understand your users. For example, when consumers book air travel, does the time that they booked a ticket have any bearing on how much money they spent? Perhaps holiday bargain seekers log on at night, while corporate travelers book flights early in the morning. What are the margins associated with each type of travel, and how do you discover the patterns of usage?

You might start by sorting through log files to determine when people started, ended, or completed a booking. You could also examine several related factors. For example, did they sort by price or by travel duration? Did they express airline preferences? Did each type of buyer prefer flights during the day or at night? How many different flight options did they consider? How many visits to your site did they make before booking, and how long did they spend contemplating their purchases?

Answering these questions requires comparing and analyzing lots of web log data that is constantly being generated. Most of that information is not very important in isolation, but when you analyze it in aggregate you can begin to see the patterns and discern important trends. Using HDFS to acquire the original data and MapReduce to process it enables you to correlate variables such as time of login, number of mouse clicks, duration of each session, and which queues or pages preceded a purchase. Then you can add this answer set to your data warehouse for additional analysis.

Example #3: Multi-Channel Marketing and Sentiment Analysis

Today's retailers must contend with a multitude of overlapping touch-points including social, digital, direct, in-store, mobile, and call center. Market leaders gain insight by analyzing transaction histories and web-behavior, as well as by concatenating data from external environments such as social media, demographics, and finance. Forward looking companies combine social media feeds, customer demographic information, psychographic data (values, attitudes, interests, or lifestyles), purchase data, and network usage data to paint a complete picture of each customer's behavior, likes, and dislikes. Harnessing this information helps retailers to understand each potential buyer as a "market of one" and to present personalized, tailored offerings to individual customers. To achieve this level of personalization, retailers must find answers hidden in massive amounts of data about

customers, spending histories, inventory, pricing, marketing campaigns, and other promotions. By analyzing this data they can better understand the factors that trigger desired behavior in various segments and channels. The data also reveals the factors that impact customer loyalty and retention, such as ease of use, value for money, and the effect of customer rewards programs. Customer churn is a major problem with retailers and the right analytic solution can help them uncover the reasons behind the churn. By examining the records about customers who have defected, you can detect patterns and then search for the early signs of those same patterns in current customers. Customer interactions can be captured, aggregated, analyzed, and correlated with other KPIs like Net Promoter Scores, to develop insights into customer behavior. For example, analyzing Twitter feeds and Facebook posts can reveal quality of service issues within specific regions or customer groups.

While traditional segmentation strategies grouped customers based on channel-specific purchase cycles, value is increasingly defined by how well a company can manage interactions across any channel including mobile, web, call center, IVR, dealers, and retail outlets. Sentiment data can tell you if a particular individual likes or doesn't like your company and product. When you combine this information with other e-business data, you can also tell if they are a big spending customer, a regular customer, or not yet a customer. You can also see if they are influencing other people in your customer database.

When you combine all this data and analyze it appropriately you can uncover hidden relationships that you would otherwise not be aware of. You can determine behavior patterns and even predict what others might do in a similar situation.

Big Data Analysis Requirements

In the previous section, Techniques for Analyzing Big Data, we discussed some of methods you can use to find meaning and discover hidden relationships in big data. Here are three significant requirements for conducting these inquiries in an expedient way:

1. Minimize data movement
2. Use existing skills
3. Attend to data security

Minimizing data movement is all about conserving computing resources. In traditional analysis scenarios, data is brought to the computer, processed, and then sent to the next destination. For example, production data might be extracted from e-business systems, transformed into a relational data type, and loaded into an operational data store structured for reporting. But as the volume of data grows, this type of ETL architecture becomes increasingly less efficient. There's just too much data to move around. It makes more sense to store and process the data in the same place.

With new data and new data sources comes the need to acquire new skills. Sometimes the existing skillset will determine where analysis can and should be done. When the requisite skills are lacking, a combination of training, hiring and new tools will address the problem. Since most organizations have more people who can analyze data using SQL than using MapReduce, it is important to be able to support both types of processing.

Data security is essential for many corporate applications. Data warehouse users are accustomed not only to carefully defined metrics and dimensions and attributes, but also to a reliable set of administration policies and security controls. These rigorous processes are often lacking with unstructured data sources and open source analysis tools. Pay attention to the security and data governance requirements of each analysis project and make sure that the tools you are using can accommodate those requirements.

Tools for Analyzing Big Data

There are five key approaches to analyzing big data and generating insight:

- *Discovery tools* are useful throughout the information lifecycle for rapid, intuitive exploration and analysis of information from any combination of structured and unstructured sources. These tools permit analysis alongside traditional BI source systems. Because there is no need for up-front modeling, users can draw new insights, come to meaningful conclusions, and make informed decisions quickly.
- *BI tools* are important for reporting, analysis and performance management, primarily with transactional data from data warehouses and production information systems. BI Tools provide comprehensive capabilities for business intelligence and performance management, including enterprise reporting, dashboards, ad-hoc analysis, scorecards, and what-if scenario analysis on an integrated, enterprise scale platform.
- *In-Database Analytics* include a variety of techniques for finding patterns and relationships in your data. Because these techniques are applied directly within the database, you eliminate data movement to and from other analytical servers, which accelerates information cycle times and reduces total cost of ownership.
- *Hadoop* is useful for pre-processing data to identity macro trends or find nuggets of information, such as out-of-range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. Organizations primarily use Hadoop as a precursor to advanced forms of analytics.
- *Decision Management* includes predictive modeling, business rules, and self-learning to take informed action based on the current context. This type of analysis enables individual recommendations across multiple channels, maximizing the value of every customer interaction. Oracle Advanced Analytics scores can be integrated to operationalize complex predictive analytic models and create real-time decision processes.

All of these approaches have a role to play uncovering hidden relationships. Traditional data discovery tools like Oracle Endeca Information Discovery, BI tools like Oracle Exalytics, and decision management tools like Oracle Real Time Decisions are covered extensively in other white papers. In this paper, we mainly focus on the integrated use of Hadoop and In-Database Analytics to process and analyze a vast field of new data.

Types of Processing and Analysis with Hadoop

Hadoop is a popular choice when you need to filter, sort, or pre-process large amounts of new data in place and distill it to generate denser data that theoretically contains more "information". Pre-processing involves filtering new data sources to make them suitable for additional analysis in a data warehouse.

For example, a concert promoter might want to analyze twitter feeds to determine how attendees liked the staging, set list, costumes, and warm-up band associated with a new Lady Gaga tour. They might begin by collecting tweets related to the artist using hash tags like "#Gaga", "#concert", "#Palladium" etc. The sentiment of each tweet can be determined by parsing the text and comparing it with positive and negative words in the English dictionary. In conjunction with MapReduce, Hadoop can process a huge amount of data in parallel on multiple servers, then re-combine it into a unified answer set or integrate it with other types of enterprise data. The resulting data set can be imported into a data warehouse for data mining and predictive analytics.

Analyzing social media from fans and concertgoers illustrates the speed at which consumer sentiments can shift online. Sports teams, elected officials and other public figures can utilize a similar strategy to identify subtle nuances in the attitudes of the general public—and respond accordingly. Any commercial organization that has a customer database can take the analysis a step further by determining how positive and negative attitudes impact total sales volume, support inquiries, and other key metrics. Pre-processing social media data with Hadoop is often the first step to predict customer behavior, anticipate cross/up-sell opportunities, improve marketing

campaign response rates, prevent churn, and analyze shopping carts to discover associations, patterns and relationships. Hadoop is also a great tool for filtering and pre-processing the data in weblog files.

Corporate data warehouses don't become obsolete in the big data world. In fact, they become more important as you discover new types of analyses and new sources of data to pre-process and feed into your existing decision support framework. Once you have sorted, summarized, and "sessionized" that data (broken it down into individual customer sessions), you are ready to load the summaries into a data warehouse for analysis. You may decide to join the sessionized information with customer purchase records from an ERP system, and then analyze the results to obtain a clearer view of which web actions lead to which types of purchases.

In-Database Processing with Oracle Advanced Analytics

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 popular open source programming language for statistical analysis. Analysts, data scientists, researchers, and academics commonly use R, leading to a growing pool of R programmers.

Once data has been loaded into Oracle Database, users can avail themselves of *Oracle Advanced Analytics* (OAA) to uncover hidden relationships in the data. Oracle Advanced Analytics, an option of Oracle Database Enterprise Edition, offers a combination of powerful in-database algorithms and open source R algorithms, accessible via SQL and R languages. It combines high-performance data mining functions with the open source R language to enable predictive analytics, data mining, text mining, statistical analysis, advanced numerical computations and interactive graphics—all inside the database.

Oracle Advanced Analytics provides all core analytic capabilities and languages on a powerful in-database architecture. These analytic capabilities include data-mining algorithms implemented in the database, native SQL functions for basic statistical techniques, and integration with open-source R for statistical programming and access to a broader set of statistical techniques.

This powerful analytic environment offers a tremendous range of capabilities to Oracle Database customers tackling big data projects by minimizing data movement and ensuring inherent security, scalability, and performance. It includes data mining tools that let you create complex models and deploy them on very large data sets. You can leverage the results of these predictive models within BI applications.

For example, you can use regression models to predict customer age based on purchasing behavior and demographic data. You can also build and apply predictive models that help you target your best customers, develop detailed customer profiles, find and prevent fraud, and solve many other analytic challenges.

Efficient Data Mining

The data mining tools in OAA enable data analysts to work directly with data inside the database, explore the data graphically, build and evaluate multiple data mining models, and deploy predictions and insights throughout the enterprise. It includes 15 data mining algorithms for classification, clustering, market basket analysis, fraud detection, and text mining that can be applied to solve a wide range of data-driven problems. It also includes a dozen algorithms that you can use to build and deploy predictive applications that automatically mine star schema data to deliver real-time results and predictions. Because the data, models and results remain in the Oracle Database, data movement is eliminated, information latency is minimized and security is maintained. Using standard SQL commands you can access high performance algorithms in the database to mine tables, views, star schemas, and transactional and unstructured data. Anyone who can access data stored in an Oracle Database can access OAA results, predictions, recommendations, and discoveries using standard reports and BI tools.

Statistical Analysis with R

Oracle Advanced Analytics has been designed to enable statisticians to use R on very large data sets. Analytic models can be written in R. The associated tables and views in Oracle Database appear as R objects. Thus there is no need to write SQL statements. Analysts can write R code to manipulate the data in the database.

By running R programs right in the database, there is no need to move data around. This integrated architecture ensures exceptional security and performance, since you can apply massive, scalable hardware resources to complex problems. OAA supports existing R scripts and third party packages as well. All existing R development skills, tools, and scripts can run transparently with OAA, and scale against data stored in Oracle Database 11g.

The tight integration between R, Oracle Database, and Hadoop enables analysts to write one R script that can run in three different environments: a laptop running open source R, Hadoop running with Oracle Big Data Connectors, and Oracle Database. It is easy to link the results of the analysis to business analytics tools such as Oracle Business Intelligence and Oracle Exalytics, as described in the following section.

Linking Hadoop and Oracle Database

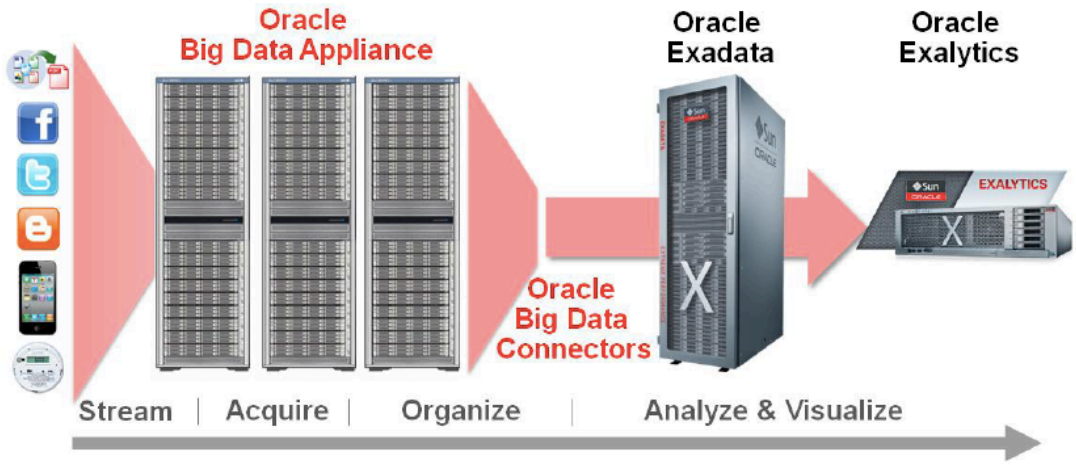
There are two different options for linking data and interim results in Hadoop with your Oracle data warehouse. Depending on your use case, you may want to load Hadoop data into the data warehouse, or leave it in place and just query it using SQL.

Oracle Loader for Hadoop provides an easy way to load HDFS data into an Oracle data warehouse. It uses MapReduce to create optimized data sets that can be efficiently loaded into Oracle Database. Unlike other Hadoop loaders, it generates Oracle internal formats, permitting it to load data faster with fewer system resources. Once loaded, the data can be accessed with traditional SQL-based Business Intelligence tools.

Oracle SQL Connector for HDFS is a high-speed connector for accessing HDFS data directly from Oracle Database, bridging the gap between HDFS and data warehouse environments. The data stored in HDFS can then be queried via SQL, joined with data stored in Oracle Database, or loaded into Oracle Database.

Oracle's Big Data Platform

Oracle has three engineered systems that solve different parts of the big data problem. Each platform includes all the necessary hardware and software necessary for extreme data processing. All components are pre-integrated and ready to deploy and operate. Oracle has done the hard work of tying these engineered systems together so that you can extract value from your data via an advanced big data platform with integrated analytics. This complete solution includes multiple systems handling data acquisition, loading, storage, management, analysis, integration and presentation so that you can quickly extract value from big data with integrated analytics.



Oracle Big Data Appliance includes a combination of open source software and specialized software developed by Oracle to address big data requirements. Residing at the front end of the big data lifecycle, it is designed to acquire and organize big data efficiently, and to be the most cost effective platform to run Hadoop. For more information on the effectiveness of this approach, see the white paper "Getting Real About Big Data: Build Versus Buy" from the Enterprise Strategy Group.

Oracle Exadata Database Machine delivers extreme performance and scalability for all types of database applications. It is the fastest platform available for running Oracle Database and the associated analytics discussed in this paper.

Oracle Exalytics is an engineered system that includes an enterprise BI platform, in-memory analytics software, and hardware optimized for large-scale analytics. With tools for advanced data visualization and exploration, it enables customers to obtain actionable insight from large amounts of data. When Oracle Exalytics is used with Oracle Advanced Analytics, customers have a comprehensive platform that delivers insight into key business subjects such as churn prediction, product recommendations, sentiment analysis, and fraud alerting.

Conclusion: Analytics for the Enterprise

Organizations in every industry are trying to make sense of the massive influx of big data, as well as to develop analytic platforms that can synthesize traditional structured data with semi-structured and unstructured sources of information. When properly captured and analyzed, big data can provide unique insights into market trends, equipment failures, buying patterns, maintenance cycles and many other business issues, lowering costs, and enabling more targeted business decisions.

To obtain value from big data, you need a cohesive set of solutions for capturing, processing, and analyzing the data, from acquiring the data and discovering new insights to making repeatable decisions and scaling the associated information systems.

Oracle Advanced Analytics is ideal for uncovering hidden relationships in big data sources. Whether you need to predict customer behavior, anticipate cross/up-sell opportunities, improve marketing campaign response rates, prevent churn, analyze "market baskets" to discover associations, patterns and relationships, leverage influencers in social networks, reduce fraud, or anticipate future demand, Oracle Advanced Analytics can help. When used in conjunction with open source tools such as Hadoop and MapReduce, this powerful analytic solution delivers everything you need to acquire, organize, analyze and maximize the value of big data within the enterprise while fulfilling fundamental requirements for minimizing data movement, leveraging existing skill sets, and ensuring high levels of security.

Big Data Analytics
March 2013

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

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

Hardware and Software, Engineered to Work Together