Technical white paper

“R” you ready?

Turning big data into big value with the HP Vertica Analytics Platform and R

Table of contents

Executive summary

Most agree that data mining can provide much-needed business value in many different ways. It can help answer questions such as:

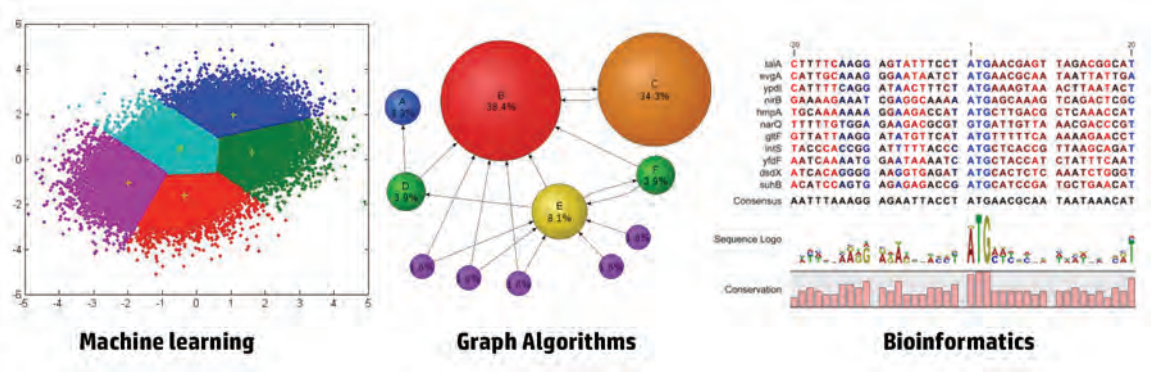
- **Financial services**—What is the probability of default for each mortgage in our portfolio?
- **Sensor data**—What is the probability of failure for each of my in-home devices?
- **Health care**—What is the probability that this medical insurance claim is fraudulent?

Yet despite this proven business value, many organizations are not taking advantage of advanced analytics with data mining. That's because of the many inhibitors to its widespread implementation. These include complexity, total cost of ownership, and questions about performance. For advanced analytics with data mining to be effective, large sample sizes are often required, and performance is critical to obtaining accurate and timely results.

Today, HP moves your organization beyond these inhibitors with the HP Vertica Analytics Platform and integration with R, a powerful language for statistical computing. The HP Vertica Analytics Platform is a proven, high-performance data analytics software platform, while R is one of the most popular open-source data mining and statistics software offerings in the market today. The combination of the two technologies helps you turn big data into big value.

The integration of R—a no-charge offering—into the HP Vertica Analytics Platform lets your enterprise sift through your data quickly to find anomalies using advanced data mining algorithms provided by R. Now, no complex import, export, or extract/transform/load (ETL) jobs are required. By integrating data mining into your processes, people in your organization are poised to make better business decisions, in less time, based on data mining results.

Figure 1: Examples of big data analytics in R



The data mining challenge

While the potential benefits are well documented, many organizations cannot achieve their expected results from data mining. The reasons for this come down to three barriers:

- **Complexity**—Data mining requires specialized domain expertise and can be difficult to integrate into applications.
- **Cost**—The initial costs of software, hardware, and implementation are a concern for many organizations.
- **Performance**—Many algorithms associated with data mining are compute-intensive, which limits the potential areas for application because decision makers cannot obtain results in a timely manner.

When considering these barriers, many organizations conclude they are not ready for data mining—it just takes too much time and effort to obtain justifiable business value.

While some business intelligence (BI) vendors do provide certain levels of data mining integration, many of them have limited their offerings to provide only a subset of data mining algorithms, typically at a large expense. As a result, many organizations have either limited their deployments of data mining to specific departments or have looked for alternatives to implementing data mining, mostly by writing custom code that is expensive to maintain and error-prone.

Data mining implementation options

You basically have three ways to implement powerful statistical tools like R:

- **Data mining and statistics with a BI tool.** Some BI tools allow for third-party tool integration to make calls to dynamic-link libraries (DLLs) or jar files, and then provide access to the statistical capabilities using the tool's expression builder. This kind of implementation typically consumes the BI server resources and negatively impacts the end-user response times. Some BI vendors also include integration with proprietary statistical software packages, but these packages often include only a small subset of statistical algorithms for a high price, and once again typically consume the same resources of the BI server, which most likely has not been optimized for statistical processing.
- **Data mining and statistics through native tools.** Some organizations have specialized statistics experts who are the only ones allowed to implement any kind of statistics. This restriction typically stems from a fear that the possibility of misuse and misunderstanding is too high to allow wider use of statistics, which are often kept siloed to a special isolated group.
- **Data mining and statistics with the database.** Some enterprises integrate their preferred statistical algorithms into their databases and then leverage the capability across their entire organizations.

The integration of R into the HP Vertica Analytics Platform follows this third approach. This white paper outlines why HP believes this is the best approach for your organization. Specifically, the paper explores how the HP Vertica implementation of R allows your organization to overcome today's data mining challenges and obtain big value from big data.

Solution overview

Starting with Version 6 of the HP Vertica Analytics Platform, HP now offers an optional no-charge download of a new package that incorporates R. This package allows your organization to implement advanced analytics using the data mining capabilities of one of the most popular open source statistics packages in the market today. Listed below is a brief summary of some of the benefits of the new HP Vertica Analytics Platform with R integration and how this combination can help your organization overcome today's data mining challenges.

Key features and benefits

The HP Vertica Analytics Platform is designed for speed, scalability, and simplicity. Among other features and benefits, the platform:

- Uses a SQL database for data mining analytics
- Runs on industry-standard x86 hardware
- Is based on a massively parallel processing (MPP) columnar architecture that scales to petabytes
- Reduces footprint via advanced data compression
- Provides extensible analytics capabilities
- Is easy to set up and use
- Provides elasticity to grow and shrink as needed
- Offers an extensive ecosystem of analytic tools

Figure 2: HP Vertica Analytics Platform

The HP Vertica Analytics Platform features a high-performance MPP, cluster-based architecture that implements a "divide and conquer" strategy applied to SQL queries. This same strategy can now be also applied to some of your advanced analytics data mining algorithms using R.

Some data mining algorithms lend themselves to a division of work that is similar to how the HP Vertica User Defined Extension (UDX) framework transform functions work, and use standard SQL-99 syntax called "windowing." In addition, the R programming model is vector-based, which in concept is very similar to the column-based architecture of the HP Vertica Analytics Platform.

Overcoming the challenges

Overcoming the complexity challenge

Analyst perspective

Familiar tools

With the R integration using standard SQL-99 syntax, your developers and end users can quickly implement new data mining algorithms into their applications, because they are already familiar with SQL. This makes using the algorithms much easier, as they are embedded within the database itself. Your developers and users can now access the algorithms using their favorite GUI and BI tools.

Proper usage

Proper implementation and applicability of data mining and statistics can be challenging without the right level of training and knowledge. Many organizations augment their teams with specialists or create a center of expertise for this focused area of need. If you take this approach, this group can then collaborate with your business domain experts to ensure the proper implementation of your new algorithms in your database. Your users can then leverage powerful statistics in their daily business analysis work and reduce time to market, because statistics can now be treated as any other standard database function.

Simplified access and streamlined process

The possibility for error is reduced because you no longer need complex ETL products to import and export the data into the statistical package. And, the entire process is streamlined so that any BI tool or ETL tool in your organization can leverage the new capabilities because they are now in your database.

Implementation perspective

Installation

A complete software package is available for installation with samples included.

Framework

User defined extensions (UDXs) let you execute business logic best suited for analytic operations that are typically difficult to perform in standard SQL. The HP Vertica Analytics Platform framework functions as an easy and powerful onramp for creative use of open source libraries and third-party software.

This framework empowers the HP Vertica Analytics Platform developer community to tackle big data problems with high-performance parallel processing and the efficiency of in-process execution. This community is active and growing with free and useful examples available at GitHub under the keyword "vertica." The implementation of R was accomplished using this existing framework by introducing a new language type variable appropriately named R. For more details, see the blog post How to implement "R" in Vertica.[1]

Overcoming the cost challenge

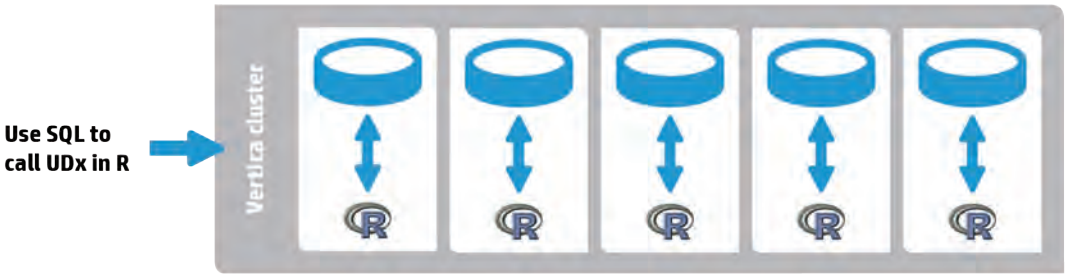
As indicated earlier, this R integration offering is free of charge. Better still, if your existing HP Vertica Analytics Platform environment is not at full capacity, you can leverage your current cluster. You don't need additional hardware or software.

Overcoming the performance challenge

The HP Vertica Analytics Platform's method of implementation, using the UDX framework mentioned above, allows you to leverage the existing Vertica cluster and realize immediate benefits:

- Move the processing close to the data, which eliminates network bottlenecks and the latency associated with the extraction, transformation, loading, and prepping of data for the data mining tool.
- Avoid making multiple copies of data. Pay for storage only once, and do so where it is most cost effective—in the HP Vertica Analytics Platform's highly available and densely encoded native storage management platform.
- Fully leverage the aggregate power of multiple nodes, multiple cores, scaled out I/O, and the cluster's memory footprint.
- Increase the ROI on your hardware investments by doing more transformation and advanced analysis in your HP Vertica database.
- Avoid extraction latency and the creation of multiple copies of data. Data can quickly become stale once it has been extracted to import into another data mining tool. This is not the case with the HP Vertica Analytics Platform.

Figure 3: The benefits of the HP Vertica Analytics and R



Key differentiators

A choice of analytical approaches

The HP Vertica Analytics Platform provides three different approaches to analyzing your data for value within your database:

- Analytics with SQL queries
- Analytics with data mining algorithms
- Analytics using external processes and custom logic

Analytics with SQL queries

HP Vertica Analytics Platform incorporates many advanced analytics that are built into the core engine. For example, some HP customers have been able to implement churn analysis without using data mining or statistics. Now they can leverage many R data mining capabilities to potentially obtain even more accurate results.

Listed below are just some of the advanced analyses that can be conducted using SQL with the core HP Vertica Analytics Platform engine:

- Time series gap interpolation
- Event window sessionization
- Social graphing
- Event series joins
- Pattern matching
- Statistical and linear regression

Analytics with data mining algorithms

The integration of R into the HP Vertica Analytics Platform enables a wide range of data analytics, including:

- Regression testing
- Statistical modeling
- Linear spatial models
- Time series forecasting
- Geo-statistical
- Text mining

For more examples on what is available with R, see the "Use cases" section.

Analytics using external processes and custom logic

User-defined functions (UDFs) provide a means to execute business logic best suited for analytic operations that are typically difficult to perform in standard SQL. Here is an example of how you can use UDFs to analyze web logs and click-stream data.

Web server log and click-stream analysis. Augmenting its strengths in high-performance in-database sessionizing, PageRank, user activity, and pattern matching, the HP Vertica Analytics Platform supports high-performance in-database web server log parsing and click-stream analysis. These capabilities are implemented as extensions to familiar SQL analytic capabilities.

With the web package, HP Vertica Analytics Platform directly imports web server logs from Microsoft® IIS and the open WC3 format (produced by Apache and many commercial products). This both eliminates the requirement for external applications and reduces the latency between online events and final analysis. The platform extracts all fields from each web server log format and executes in parallel on multiple nodes and cores.

You can leverage the dense encoding and compression of the HP Vertica Analytics Platform to retain the detailed history needed for detailed profiling. Results from deep analysis of trends and ranking, as well as identification of high-value customer segments, can feed directly into models used by dynamic content applications for a personalized and highly engaging web experience.

A data-centric approach to analytics

While various database vendors can claim R integration, HP Vertica Analytics Platform's implementation can provide better value to your organization because of its simplicity and performance. That's because other vendors typically take an R-centric approach, which means users have to know and use the R suite of programming tools.

HP Vertica Analytics Platform's implementation, in contrast, is a more data-centric approach, which shields your users from having to know and use the R language. This means your users can continue to use their favorite BI or query tools and now have access to R capabilities.

As an example, once the HP Vertica Analytics Platform implements the R K-Means function, it can be used by anyone who needs to group subjects together, using a sophisticated data mining clustering technology across many business domains.

To invoke the new R function, you can use standard SQL syntax, such as:

```
select kmeans(geonameid, latitude, longitude) over () from geotab_kmeans;
```

The example above shows how you would invoke the R function with a "points in space" or location-related scenario. The following example shows how to use it in a Moneyball example. For more detail, see the blog post "Vertica Moneyball and 'R'. The perfect team!"[2]

```
select kmeans(playerid, WHIP, IPOUTS) over () from bestpitchers;
```

A complement to BI systems

Some BI vendors are now implementing R into their products. As a result, you can leverage their graphics capabilities to build dashboards showing data mining results as well as the high-performing HP Vertica Analytics Platform database to get results faster.

In addition, you can have the BI engine "push down" the execution of BI algorithms to the HP Vertica Analytics Platform database, leveraging the power of the database to do the "heavy lifting" of this processing. This offloads the BI server of this workload, allowing your organization to partition the work to the best and most powerful machines available in your network—your HP Vertica Analytics Platform database.

Use cases

The integration of R into the HP Vertica Analytics Platform supports a wide range of data analytics use cases, including:

- Behavior analytics—The analysis of an individual's behavior, such as buying behavior, and making meaningful predictions based on past and current data.
- Click-stream analyses—The recording of the parts of the screen a computer user clicks on while web browsing or using another software application. Click-stream analysis is useful for web activity analysis, software testing, market research, and analyzing employee productivity.
- Network analyses—The analysis of network information and the relationships between various nodes in the network. Examples include social networks and computer networks.
- Customer analytics—A process by which data from customer behavior helps make key business decisions through market segmentation and predictive analytics.
- Compliance testing—To determine whether a product or system meets some specified standard that has been developed for efficiency or interoperability.
- Loyalty analysis—A type of customer analytics focusing on a person's commitment to a product, company, or brand.
- Campaign management—To provide advanced management capabilities for the data used to conduct outbound marketing campaigns.
- Promotional testing—Using data typically associated with marketing and campaign management systems to identify the best criteria to be used for a particular marketing offer.
- Claims analyses—The analysis of claim data, such as insurance or warranty claims, to typically identify anomalies such as fraud or identify product defects early in the product release phase.
- Patient records analyses—The analysis of medical records associated to patients to identify patterns to be used for improved medical treatment.
- Clinical data analyses—The analysis of clinical data and its impact on patients to identify patterns to be used for improved medical treatment.
- Fraud monitoring—An intentional deception made for personal gain or to damage another individual. Monitoring is the process of identifying and predicting this activity.
- Financial tracking—Typically deals with ensuring regulatory and compliance with financial-related data.
- Tick data back-testing—Analyzing tick-by-tick historical market data identifying patterns compared to historical records.

An industry scenario: monetizing telecommunications data

Telecommunication companies (telcos), particularly those that provide multiple services to the customer (voice, data, SMS, mobile), often understand customer behavior but historically could not monetize this valuable asset. Telcos can now find new ways to monetize their subscriber and network data over mobile, local, and social channels.

Mobile channel

Imagine that you, as a mobile advertiser, can know whether a user is a sports or fashion fan, or is looking for a restaurant to eat lunch at in the next 30 minutes, or is shopping for a new car. A telco can determine this data about its subscribers (opt-in, of course) using deep packet inspection techniques to extract click-stream information from network traffic.

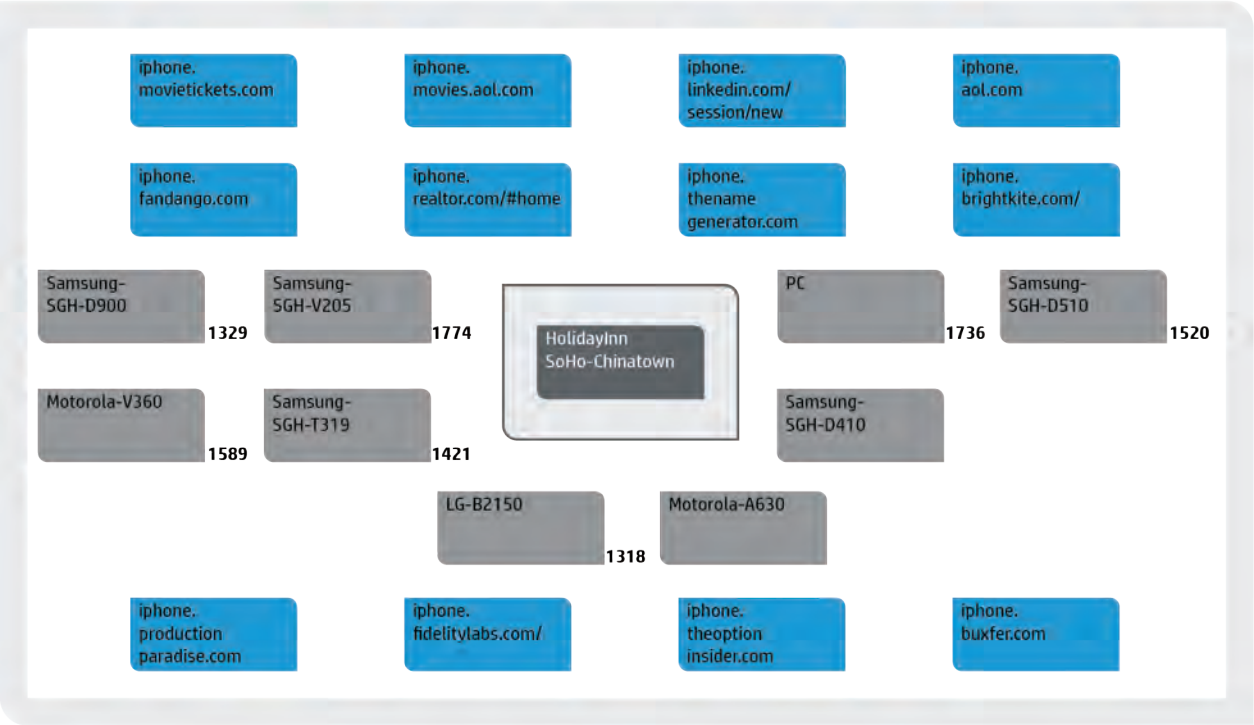
You can further categorize historical customer browsing behavior across multiple websites and browsing endpoints—such as desktop, mobile, or TV—into broad categories of interests. You can use these categories to build a rich profile that includes long-term and short-term interests. You can then apply a statistical model to the data to predict a customer's probability of clicking on an ad on a certain topic in this hour, day, or month.

A telco can use this model for its own promotions. For example, if you often watch movies on your phone, you may be interested in a customized video data plan with better quality of service. Or the telco can offer this model as a service to mobile advertisers to improve the targeting of ads for individual mobile users.

Local channel

A telco also has location data about subscribers and their calling behaviors. To leverage this data, you can use a K-Means clustering technique to segregate customers based on from where they are most likely to make a phone call. You can also use this information to determine whether a particular subscriber may experience worse quality of service due to poor cell coverage in the subscriber's preferred location. Or you can determine how many subscribers use a particular calling location and use that information for capacity planning. Figure 4 shows the relationships among the cell tower location, subscriber mobile devices, and routers.

Figure 4: Wireless network relationships



Social channel

A telco also has a compelling social graph by knowing the calling behavior of its users. You can calculate a social graph of friends and family for a subscriber base of millions of users with hundreds of millions of edges.

You can also use the PageRank social graph analysis technique (similar to what Google uses to rank search results) to identify a list of "influencers." When targeting a new telco service, you can focus first on a smaller influencer set and rely on its members to pass the word on to others.

Tying it all back together, once you have the set of influencers, you can use their personalized profiles to provide personalized incentives if they spread the word—for example, giving baseball tickets to sports fans and a free month of video on demand to movie buffs.

Key takeaways

Accelerate time to market

As always, time to market is extremely important. The HP Vertica Analytics Platform's implementation of R is designed to help you accelerate time to market for data analytics. From the outset, this integration saves you implementation time, because you can leverage UDX integration.

Your users, in turn, can leverage standard SQL syntax, so they don't have to become experts in R. And from a performance perspective, you can capitalize on the parallelism of the HP Vertica Analytics Platform's multi-node architecture to accelerate system performance and time to results.

Provide a comprehensive analytical solution

With this integration, the number of possible data mining and statistical algorithms is almost endless because of the constant innovation and openness of the R project. You can use many sophisticated data mining algorithms, including:

- K-Means clustering to segment customers based on geography, usage patterns, or other characteristics
- PageRank to identify the influencers among your customers and users
- K-Nearest Neighbors classification to identify products or users that are similar to other products or users based on proximity in a feature space
- Naïve Bayes classification to group events into categories based on historical data
- In-database scoring—sometimes referred to as in-database analytics—for the integration of data analytics into the database for situations such as deriving a score for an individual

Another benefit of using the HP Vertica Analytics Platform architecture is the opportunity to implement "Ensemble," which allows you to combine multiple algorithms into one function and then have the function pick the most accurate one for results. Your options are virtually endless.

Control your costs

Integrating R into your HP Vertica Analytics Platform can be a cost-effective way to implement data mining. If your existing HP Vertica Analytics Platform environment is not at full capacity, you can leverage your current cluster. You don't need additional hardware or software.

Meet service level agreements

Some big data problems demand better use of the hardware architecture to deliver timely results. For example, K-Means is a powerful but compute-intensive algorithm that can involve multiple iterations to increase the accuracy of the results. The HP Vertica Analytics Platform implementation of R takes advantage of clustering and parallelism to help you improve the accuracy of results with this algorithm while meeting your service level agreements.

To learn more

To test drive HP Vertica Enterprise Edition, visit **vertica.com/evaluate**.

HP also offers HP Vertica Community Edition software, a free version of HP Vertica Enterprise Edition limited to one terabyte of data and three nodes. Sign up for HP Vertica Community Edition at **vertica.com/community**.

[1] http://www.vertica.com/2012/10/02/how-to-implement-r-in-vertica/
[2] http://www.vertica.com/2012/09/11/vertica-moneyball-and-r-the-perfect-tea

---

Get connected

hp.com/go/getconnected

Current HP driver, support, and security alerts
delivered directly to your desktop

4AA4-4207ENW, Created October 2012