White Paper
Analytics and Big Data

From Hindsight to Insight to Foresight

Extend Analytical Capabilities with Vertica

Vertica is a powerful database platform that provides an ideal environment for in-database SQL analytics functions as well as in-database machine learning for predictive analytics. Tight integration with R and other advanced analytics libraries make Vertica extensible so that almost any type of analytics can be achieved. This white paper explores the wide range of analytical functions in the platform—including standard SQL-99 conventions, value-added analytics, in-database machine learning, advanced analytics using custom logic, and user-defined extensions.

The paper incorporates brief case studies that summarize real-world applications of Vertica. In addition, the paper explains how you can implement the built-in capabilities of Vertica and develop your own next-generation big data analytics functions using the platform's C++, Java, and R SDKs.

Table of Contents

The power of big data analytics platform enables the democratization of data that can help your organization make better-informed decisions, compete more effectively, and gain a real return on information.

Gaining an Edge in a Data-Driven World

In today's business environment, competitive advantage increasingly hinges on how well organizations can turn mountains of data into meaningful insights for customer behavior, market opportunities, business trends, product quality, security threats, and more. This reality of a data-driven world has prompted many organizations to adopt big data analytics platforms to gain actionable insights from the data.

In many cases, the drive to create or enhance an analytics program begins with a single project that is being created from the ground up or that is not generating the information needed for the organization. That's when many forward-thinking organizations turn to big data analytics software to push the right information at the right time further into their organizations.

As your organization considers data analytics platforms for projects, it's important to keep an eye on the larger picture. In particular, in developing an overall strategy, you need to consider the capabilities to ingest big data, perform and scale analytic queries handling future data volumes, prepare for more widespread use of analytics by end users, and even prepare for future types of advanced analytics.

Ultimately, the power of big data analytics platform enables the democratization of data that can help your organization make better-informed decisions, compete more effectively, and gain a real return on information. These ideas are at the heart of the analytics-driven enterprise.

Variations in the Depth of Analytics

When designing your analytical platform, keep in mind that they vary greatly in the depth of analytics offered. NoSQL databases, for example, often offer only rudimentary implementations of SQL on Hadoop. Some cloud-based SQL analytics solutions offer only a subset of the analytics that may be needed to gain a complete understanding of your business. For the business, this reveals itself in the speed and agility at which data analytics can be delivered to information consumers.

Another common tool that is sometime improperly selected for analytics are Hadoop-based query engines. Query engines like Presto, Hive and Impala offer rudimentary SQL. However, they lack in providing production-based features that let them handle concurrent queries, for example. Custom code is often needed to achieve business goals and production-ready status. In this scenario, both development of queries and movement of the data make advanced analytics too burdensome. Business and analytical innovation suffers at the hand of onerous coding, lack of concurrency and difficult data transformation. So what does it take to create an analytics-driven enterprise? Many organizations begin with the core capabilities of Vertica.

Analytical functions in Vertica span from standard SQL-99 conventions to value-added SQL capabilities to in-database machine learning.

Creating an Analytics-Driven Enterprise

In the analytics-driven enterprise, big data is leveraged to help decision makers understand the past—what happened and why it happened—and to gain the insight and foresight needed to make better predictions and decisions about the future, as shown in Figure 1.

So what does it take to create an analytics-driven enterprise? Many organizations begin with the core capabilities of Vertica. Let's explore these capabilities.

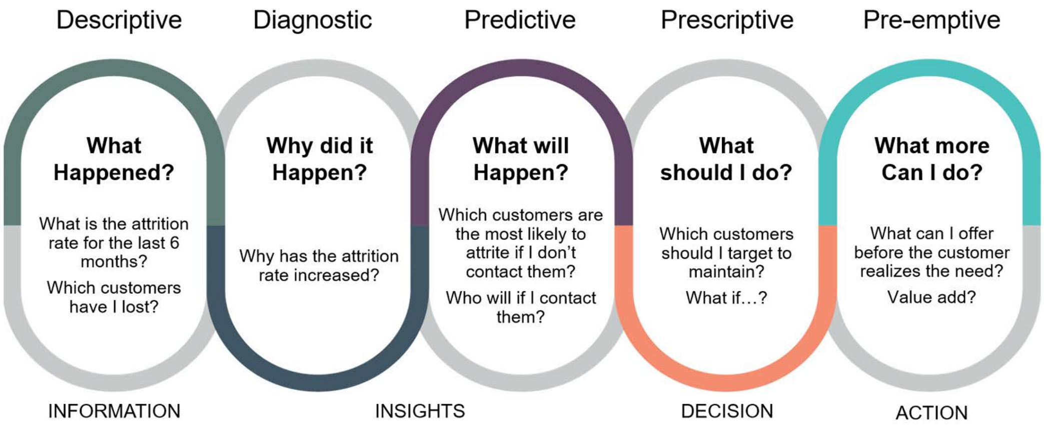
Descriptive
Diagnostic
Predictive
Prescriptive
Pre-emptive
What Happened?
What is the attrition rate for the last 6 months?
Which customers have I lost?
Why did it Happen?
Why has the attrition rate increased?
What will Happen?
Which customers are the most likely to attrite if I don't contact them?
Who will if I contact them?
What should I do?
Which customers should I target to maintain?
What if…?
What more Can I do?
What can I offer before the customer realizes the need?
Value add?
INFORMATION
INSIGHTS
DECISION
ACTION

Figure 1. From hindsight to insight to foresight—this sequence illustrates the ways an analytics-driven enterprise can use big data analytics to improve customer retention.

Analytical Functions in Vertica

The analytical functions in Vertica span from standard SQL-99 conventions to value-added SQL capabilities to in-database machine learning. These analytics functions perform and scale well for big data needs by leveraging Vertica core capabilities of column storage, advanced compression, and massively parallel processing (MPP) architecture.

The easy-to-use SQL analytics functions in Vertica greatly simplify the writing of complex queries.

By combining the scalability features with the analytical and machine learning features, you can use the Vertica engine to training the machine learning models without having to move data. You can avoid the common practice of moving data, or subsets of data, from the data warehouse into Apache Spark in order to train your machine learning models. Vertica scales so that you can use more data and get better precision on your models.

In addition, Vertica is built to be extensible. It allows the development of custom SQL analytics functions that can leverage the Vertica MPP architecture and data locality computation. Vertica supports popular languages—C++, Java, and R—to accelerate the development of user-defined extensions (UDx).

Vertica SQL: Standard SQL-99 Conventions

Vertica incorporates a wide range of SQL functions based on the ANSI 99 standard.

Aggregate Functions

Aggregate functions summarize data over groups of rows from a query result set. Vertica supports a wide range of aggregate functions, including:

- Simple functions like Sum, Min, Max, counts, and average
- Statistical functions to compute variance, covariance, standard deviation, and correlation
- Regular expressions and fast-performing approximate functions with user-specified error tolerance

These aggregate functions scale and perform many times better than traditional databases using Vertica core capabilities.

Analytic Functions

The easy-to-use SQL analytics functions in Vertica greatly simplify the writing of complex queries. These functions enable you to meet complex analysis requirements and reporting tasks with a few lines of standard SQL code and without writing thousands of lines of code in NoSQL frameworks.

For example, with a single SQL analytic function, Vertica can provide the following complex analytic requirements:

- Rank the longest-standing customers in a particular state
- Calculate the moving average of retail volume over a specified time

- Find the highest score among all students in the same grade
- Compare the current sales bonus each salesperson received against his or her previous bonus

Applications of analytics functions include graph analysis, triangle counting, and Monte Carlo methods. These analytical operations are not feasible in traditional databases because they don't perform well. With its columnar MPP database, Vertica performs and scales well for these complex operations.

Graph Analysis

Graph analysis can be useful for understanding relationships, like those that exist on social networks. It also can be used to detect fraud and spammers on networks.

Graph analysis has traditionally been a showstopper for relational databases. While the table structure is simple—typically just a few columns—to express and analyze the graph requires self-joins to express the connections in the graph, which causes an explosion in the number of rows in the result set due to combinatorial factors. Row-oriented database platforms are simply unable to cope with the massive volume of data created during the execution of graph queries.

Vertica, on the other hand, handles this work quite well. Its columnar features deliver extremely high levels of performance on the graph data, and its MPP architecture allows the system to scale as needed.

Triangle Counting

Triangle counting is an example of how Vertica analytics functions can solve a key aspect of graph analytics. Let's use friendship as an example: if two of your friends are also friends with each other, then the three of you form a friendship triangle. Vertica allows you to use simple SQL functions to count triangles for performance graph analysis.

Monte Carlo

The main ingredient in any Monte Carlo method is, as the name suggests, some process akin to going to the casino and repeatedly throwing the dice to see if the odds are in your favor or stacked toward the house. While this description is simple, the true beauty lies in the application of brute force; instead of applying a complex analysis to the rules and physics of a game, just go play it a number of times and see what happens.

> **"The simplicity and functionality of Vertica allowed us to rapidly develop and launch our data warehouse. This saved us time, resources and money instantly. Vertica provides data analytics for multiple departments of our company and serves as the central repository of historical data and information. Its ease of use and cost-scaling capability allows us to expand and explore new useful sources of data and convert them into useful information."**
>
> **IT ARCHITECT**
> Medium Enterprise
> Telecommunications
> Services Company

Vertica is a good SQL analytics platform for implementing Monte Carlo techniques. Conversely, the Monte Carlo method is a great tool to have on hand for analyzing data, particularly in situations where brute force seems more appealing than over-thinking things.

Geospatial

Vertica includes built-in functions for geospatial analysis. Mathematical functions, automatically installed with Vertica, let you perform common geospatial operations. Vertica provides geospatial functionality to support collection and analysis of geospatial information to add the aspect of "where."

Your organization can deliver personalized marketing with geo-precise target advertisements by combining big data with location data. You can target the most relevant customers by combining location with brand affinities, consumer passion points, and dynamic mobile behavior.

Your organization can deliver personalized marketing with geo-precise target advertisements by combining big data with location data.

Vertica supports a data loader for ESRI shapefiles, data representation in the Open Geospatial Consortium's (OGC) Well-Known Text (WKT) and Well-Known Binary (WKB) formats, and many OGC-based SQL functions for computation on two-dimensional planar data with select support for spherical functions as well.

Vertica Extended-SQL: Value-Added Analytics with SQL

Vertica offers a wide range of extended SQL analytics functions. These include capabilities for sessionization, time series analytics, event-based windows, event series joins, social media, and pattern matching.

Sessionization

Sessionization, a special case of event-based windows, is a feature often used to analyze click streams, such as identifying Web browsing sessions from recorded Web clicks during a specific period of time. Unlike brute force procedural methods that can achieve this, the approach taken in Vertica is simple, efficient, and massively parallel so that Web sessionization can be done in an ad hoc manner with various window parameters determined on the fly.

Suppose, for example, that 30 seconds may not be an average Web visit session. Vertica can automatically analyze the intervals of sessions from equal IP addresses to determine what the average session time truly is and then tokenize or sessionize the data automatically based on that parameter.

Time Series

Time series analytics evaluate the values of a given set of variables over time and group those values into a window for analysis and aggregation. Vertica is purpose-built for time series analytics. This is due to both the optimized structure and the analytical capabilities of the platform. A columnar orientation allows time series data to be sorted, compressed, and partitioned to enable optimal performance.

Vertica provides gap-filling functionality, which fills in missing data points, as an interpolation scheme. This is a method of constructing new data points within the range of a discrete set of known data points. The platform interpolates the non-time series columns in the data (such as analytic function results computed over time slices) and adds the missing data points to the output.

Vertica provides gap-filling functionality, which fills in missing data points, as an interpolation scheme.

Event-Based Windows

Event-based windows functions are a Vertica extension to the standard SQL analytics. These functions simplify the detection of events in time series data. Event-based windows let you break time series data into windows that flag on significant events within the data. This is especially relevant in financial data where analysis often focuses on specific events as triggers to other activity. The event-based window functions in Vertica assign to each input row an integer value representing the event ID, starting from 0. The event ID is incremented whenever a user-specified event happens in time series data.

For example, given an input stream of stock quotes, the stock analyst may want to place the input quotes into a new group whenever the spread (the difference between the ask price and the bid price) exceeds five cents. If we view each such group as a window of events, then the window endpoints are defined by the occurrence of certain event types.

There are two event-based window functions in Vertica—Conditional Change Event and Conditional True Event. These functions are a Vertica extension.

Event Series Joins

Vertica supports typical data warehousing query joins. The platform also provides the interpolate predicate, which allows for a special type of join. The event series join is a Vertica SQL extension that lets you analyze two event series when their measurement intervals don't align precisely—such as when timestamps don't match. These joins provide a natural and efficient way to query misaligned event data directly, rather than having to normalize the series to the same measurement interval.

Vertica natively supports path and pattern analysis through an event series pattern matching extension.

Social Media

The analytical capabilities in Vertica allow you to measure social media conversions, referrals, landing pages, and data sources. The platform's underlying optimized storage structure and sorting allows large volumes of data to be joined for efficient conversion rate analysis from social sentiment and consumer buying patterns.

Vertica's time series analytics, sessionization, and event series pattern matching extensions allow you to measure page views, duration, visits, and landing page heuristics of referring social sites and social sentiment over time.

Vertica can help you with entity extraction and sentiment analysis by automatically analyzing short text to help you understand what your community is talking about and how it feels about those topics. Vertica social media functions can be executed through a single line of SQL. You can then extract the aspects (called attributes) of a brand, product, service, or event that your users and customers are talking about. In addition, it enables you to assign a sentiment score to each of these attributes, so that you can track your community's perception on the aspects of your business that your community cares about.

Pattern Matching

Vertica natively supports path and pattern analysis through an event series pattern matching extension. The SQL MATCH extension lets you screen large amounts of historical data in search of event patterns. You specify a pattern as a regular expression and can then search for the pattern within a sequence of input events. MATCH provides subclauses for analytic data partitioning and ordering, and the pattern matching occurs on a contiguous set of rows.

Pattern matching is particularly useful for clickstream analysis where you might want to identify users' actions based on their Web-browsing behavior (page clicks).

Vertica also offers native support for funnel analysis with standard SQL analytics and analytical extensions purpose built for funnel analysis. The platform's event series pattern matching extension can be used to calculate standard conversion rates, drop-off and bounce rates, and conversion paths. You can combine standard SQL analytics, such as the lag and regular expression analytics, to analyze conversion paths.

Vertica User-Defined Extensions

User-defined extensions (UDxs) in Vertica let you execute business logic best suited for analytic operations that are typically difficult to perform in standard SQL.

Vertica's broad array of user-defined capabilities (functions, transforms, aggregates, analytics, and loading) brings the power and flexibility of procedural code closer to the data—be it structured, semi-structured, or unstructured—fully leveraging the parallel compute environment of Vertica. User-defined extensions run in process for maximum efficiency, or fenced for additional control. In either mode, Vertica's user interface makes it easy to deploy and use procedural extensions, encouraging maintainable operational practices and promoting code reuse.

Vertica's user interface makes it easy to deploy and use procedural extensions, encouraging maintainable operational practices and promoting code reuse.

Procedural vs. Declarative

For many people, it is easier to approach a solution procedurally, thinking of the steps required to break the task into a sequence of actions. By contrast, SQL is a declarative language, where the operator states the required outcome, and the database develops an optimal procedure and efficiently computes the answer.

The Vertica Optimizer is no exception—it understands how the data is distributed about the cluster, the most cost-effective join order, and generates highly efficient query plans (the steps in a procedure). Nevertheless, some operations are difficult or tedious to express in SQL, especially when the inputs are unstructured or semi-structured. For cases where procedural language is more natural, user-defined extensions are the natural choice.

Predictive Analytics

Predictive analytics is more proactive and deeper than traditional descriptive analytics. Descriptive analytics does a good job of slicing and dicing data to help answer questions on what happened or what is happening, and perhaps why did it happen. With predictive analytics, however, you can predict what will happen by learning patterns from historical data.

Machine learning is most effective when applied to very large data sets and thus is a natural fit for Vertica which is designed for fast processing of big data. Built into Vertica's core—with no need to download and install separate packages—in-database machine learning supports the entire predictive analytics process with massively parallel processing and a familiar SQL interface, allowing data scientists and analysts to embrace the power of big data and accelerate business outcomes with no limits and no compromises.

Some of the major use cases for machine learning applications revolve around classification, clustering, and prediction. Vertica's built-in machine learning algorithms cover all of these areas with algorithms like K-means, linear regression, SVM, logistic regression, and Naïve Bayes. But it's not just about the algorithms. There are two huge benefits to using Vertica for machine learning: 1) You don't have to move a smaller subset of the data out of your data warehouse for training and analytics, and 2) You gain access to all of the data preparation and ELT features of Vertica via SQL to ensure data quality.

If you want to go beyond the in-database algorithms, Vertica natively supports R for statistical computation. There are many public R libraries and functions for modeling and classification that can be easily installed on Vertica. The platform's R implementation runs in-database and allows R functions to be parallelized to take advantage of Vertica's underlying MPP platform.

Vertica's SDKs allow for the development of R functions and the integration of the wealth of public R functionality. Regarding presentation of data, many different commercial and open source visualization tools can be used with Vertica.

> **"Vertica is quite simply the 'fast MySQL' for data analysis. Powerful partitioning SQL commands have allowed us to do some clever behavioral analytics."**
>
> **ENGINEER**
> Medium Enterprise
> Computer Software Company

Data Ingestion

Vertica enables integration with third-party and open-source databases, data management tools, and data analytics packages. Depending on your preference or development skills, you can use Vertica SDKs to write custom C++, Java, or R code. These capabilities enable Vertica to simplify and scale for big data needs.

Connection options include:

- **Open Database Connectivity (ODBC)**—is a standard API for access to database management systems.
- **Java Database Connectivity (JDBC)**—is a call-level API that provides connectivity between Java programs and data sources.
- **Apache Hadoop**—is an open-source software framework for storing and processing large data sets on clusters of industry-standard hardware. Vertica natively reads parquet and ORC files (a common HDFS standard).
- **Apache Hive**—is a data warehouse package for querying and managing large data sets that reside in distributed storage.
- **Vertica Flex Zone**—is an offering that gives you the power to quickly and easily load, explore, analyze, and monetize rapidly growing forms of semi-structured and structured data.

Realizing the Benefits: Case Studies

Customer Analytics

An online gaming company with more than 100 games on various social media platforms had been experiencing explosive growth, and its analytics program was suffering from the success. As a data-driven organization, this company could no longer run its A/B tests and get reports on its tens of millions of users due to the performance of its MySQL database.

]After testing Vertica in a trial, the company implemented the platform into its production environment. Since the transition to Vertica for testing and analysis, A/B test result queries that used to take nine hours now return in 12 seconds. In all, the company adds about 800 million rows of event data per day. Its largest table is the events table with 1.2 trillion rows, and it has 50 million rows in the dimension table for users and some other smaller tables.

Operations Analytics

A large telecommunications company saved millions of dollars in equipment purchases due to a richer view of its network utilization using Vertica for network performance data analysis.

The company deployed several clusters of Vertica, empowering its analytics team to query massive amounts of data using familiar SQL commands and obtain results in a fraction of the time. The company stored and analyzed data network usage data over extended periods, making trend forecasting simple.

In addition, because Vertica seamlessly connects with the company's favorite visualization tools, the company created dashboards that showed the status of a number of different metrics based on time of day, peak hours, and over a set number of days. And analyzing causal relationships and running "what if" scenarios across billions of records using Vertica has changed the way the company looks at what's possible with data analytics.

Patient Analytics

A leading healthcare solutions company has an application platform that provides electronic health records for healthcare providers, and also helps those providers optimize processes to speed the delivery of care and eliminate waste and error.

To ensure the platform provides the rapid response and overall performance demanded by customers, the company has built into the platform some 2000 response timers. These timers detect how long certain application functions take—such as adding or accessing patient information, or entering an order for medication or a procedure. As the amount of data continued to grow, the company's existing legacy data warehouse ran into severe performance issues, and the company knew it needed to make a change.

To help you make the most of the wide-ranging capabilities of Vertica and complementary third-party solutions, Micro Focus offers a range of professional services and data scientist expertise.

After testing five other advanced analytics platforms, the company chose Vertica to help with its application platform and another healthcare solution provided by the company. With Vertica, the company integrates advanced performance monitoring and healthcare data analytics into its application platform and its proprietary healthcare information reporting solution.

The application architecture places patient information at the clinician's fingertips at the optimal time and place with appropriate metrics and reporting capabilities. Vertica delivers performance information that allows the company to stay ahead of performance issues and to analyze usage patterns for continued improvements.

Sensor Data Analytics

With the release of ProLiant Gen8 servers, Hewlett Packard Enterprise introduced a "call-home" feature that sends telemetry data back to HPE. Hewlett Packard Enterprise loaded the incoming telemetry streams and began to analyze the data for consistent patterns in configurations in hardware and software that caused challenges (i.e., warranty events). Warranty events cost Hewlett Packard Enterprise in both support costs and customer satisfaction.

Hewlett Packard Enterprise was looking to get proactive with its notifications to customers around the issues they found through analysis. For example, if you have HPE ProLiant XYZ and you install ABZ software, x percent of the time we would expect a failure. Here is how you resolve the issue.

Telemetry streams data 24x7. Every time there is a keystroke or configuration change, a log is sent back to Hewlett Packard Enterprise. HPE needs an engineer to determine the fault scenarios and fixes, but with Vertica, the engineers can see patterns where failures are happening on a micro level and fix them before they impact a large portion of the population.

Vertica provides added value in its ability to create dynamic perspectives on the fly, its speed and compression, and its support for in-database algorithms that help detect patterns.

Bring It All Together with Vertica Expertise and Support

You're not alone in your quest to become an analytics-driven organization. To help you make the most of the wide-ranging capabilities of Vertica and complementary third-party solutions, Micro Focus® offers a range of professional services and data scientist expertise.

These analytics experts can help you identify the right mix of capabilities from Vertica and integrate your analytics environment with third-party and open-source solutions using the platform's C++, Java, and R SDKs. They can help you understand best practices for analytics deployments and the approaches taken by different organizations.

Vertica enables your organization to make better-informed decisions, compete more effectively, and gain a real return on information. These ideas are at the heart of the analytics-driven enterprise.

A Growing Ecosystem

In addition to leveraging the expertise of Micro Focus, you can capitalize on the collective knowledge of the Vertica user community via my.vertica.com. This portal provides access to software downloads, documentation, and discussion forums that enable your in-house team to augment, share, and grow your own domain expertise.

Key Takeaways

Vertica is a new-age data analytics platform designed from the ground up for business analytics at the scale of big data. The platform provides an ideal environment for in-database analytics functions, as well as tight integration with R and other advanced analytics libraries.

The wide-ranging analytical functions in Vertica include standard SQL-99 conventions, value-added analytics with SQL, in-database machine learning, user-defined extensions, and big data advanced analytics using custom logic. In addition to leveraging the platform's built-in capabilities, you can develop your own next-generation analytics functions using the platform's C++, Java, and R SDKs.

Ultimately, Vertica enables your organization to make better-informed decisions, compete more effectively, and gain a real return on information. These ideas are at the heart of the analytics-driven enterprise.

Learn More At
www.vertica.com

Additional contact information and office locations:
www.vertica.com

www.vertica.com