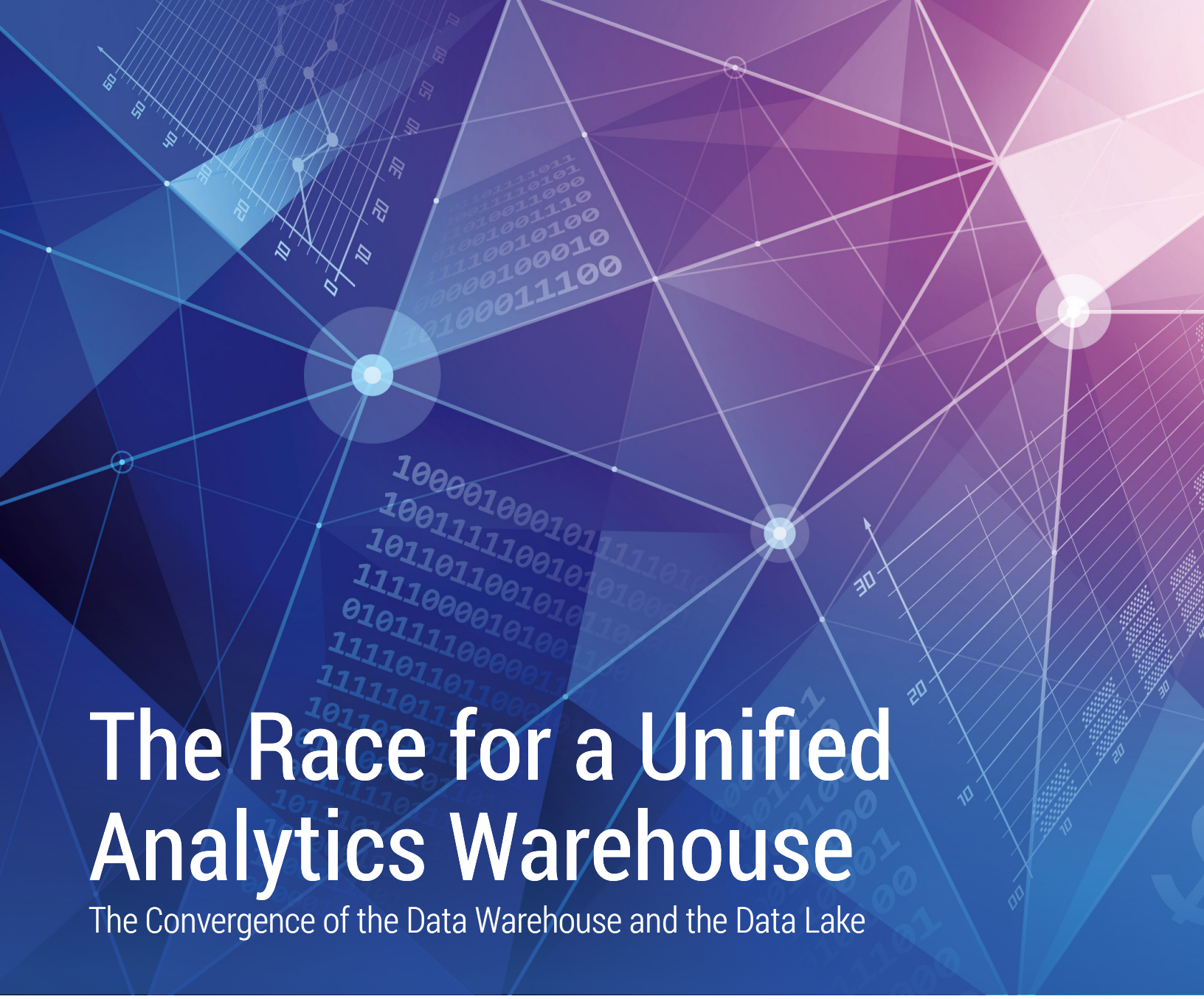


The Race for a Unified Analytics Warehouse

The Convergence of the Data Warehouse and the Data Lake

An ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) White Paper

Prepared for Vertica

By John Santaferraro

June 2020

THE RACE IS ON

The race for a unified analytics warehouse is on. The data warehouse has been around for almost three decades. Shortly after big data platforms were introduced in the late 2000s, there was talk that the data warehouse was dead—but it never went away. When big data platform vendors realized that the data warehouse was here to stay, they started building databases on top of their file system and conceptualizing a data lake that would replace the data warehouse. It never did.

Data lake platforms quickly demonstrated the value of semi-structured data. As data lakes grew in popularity, data warehouse vendors saw them encroaching on their territory, so they began to shift their databases to run on file systems and started finding ways to process semi-structured data.

Pressured by customer demands to run analytics across both the data lake and the data warehouse, vendors on both sides began working toward a more complete integration of a warehouse and lake.

Within a few years, nearly every organization that ran a data warehouse also stood up a data lake. The two existed side by side. Initially, there was some data sharing between the two platforms, but not much more. Pressured by customer demands to run analytics across both the data lake and the data warehouse, vendors on both sides began working toward a more complete integration of a warehouse and lake.The race began.

The Reason for the Race

Digital, mobile, and the Internet of Things (IoT) changed the way modern companies operated. Digital transformation proved that semi-structured data is as important as transactional, structured data. Both need to be analyzed to create a competitive advantage. Unfortunately, neither the data lake nor the data warehouse was adequate to handle the analysis of both data types.

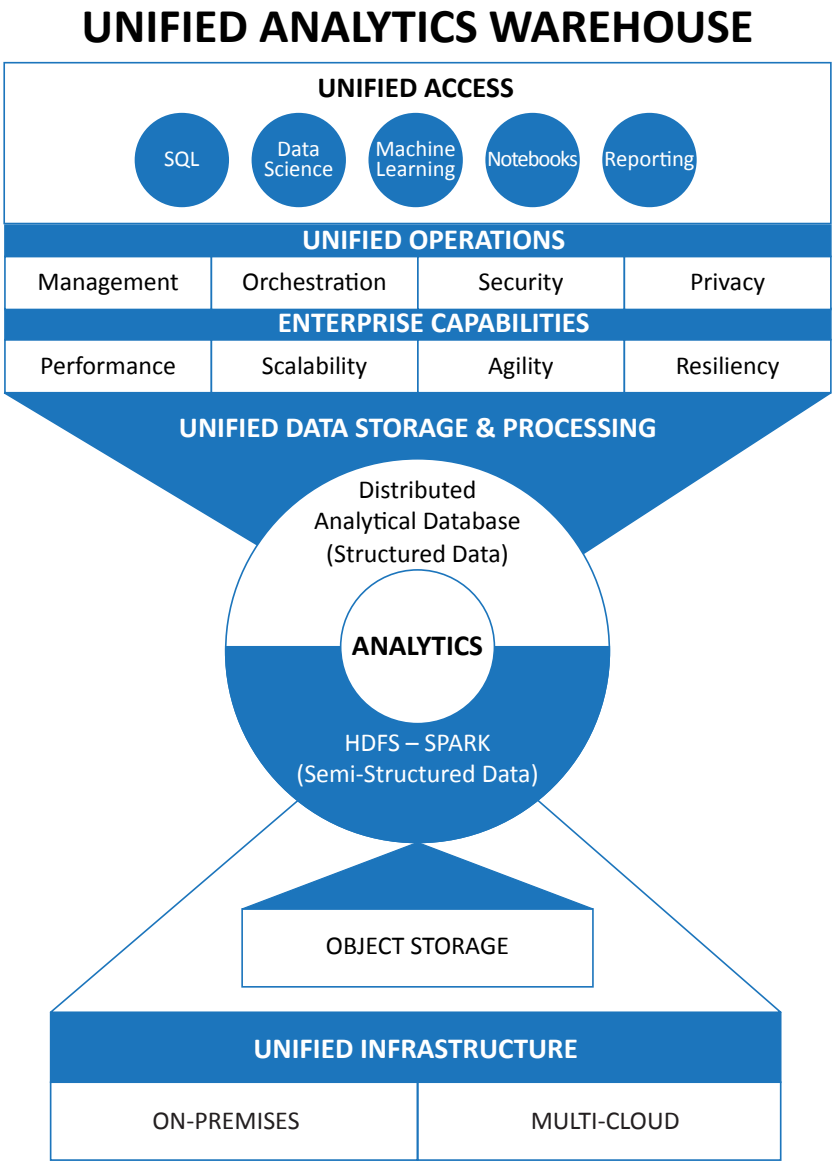
The data warehouse was inadequate because it was built for structured data and required structuring and ingestion of semi-structured data to analyze it.The data lake was inadequate because the associated database technology lacked enterprise capabilities and did not perform well. Analyzing each type of data separately was not the answer. Only structured and semi-structured data combined could deliver on the promise of complete business insights.

The Unified Analytics Warehouse

Driven by customer requirements, both camps have been consistently pushing toward a unified analytics warehouse (UAW). It is unified because it adequately handles multi-structured data in a single platform. It is an analytics platform because the primary use case for both the data lake and the data warehouse has always been analytics. The data lake has focused more on data science use cases and the data warehouse has focused more on enterprise analytics. It is a warehouse because it stores multi-structured data in an organized and accessible manner.

From the data lake side of the race, nascent data platform vendors who built their data architectures to handle semi-structured data are working to build database technology, including structured columnar storage formats and SQL query engines, on top of their file-based data storage systems. Because the data is typically stored in a file system, providing access to object storage, like Amazon S3, was either built-in or easy to build. Data lake technology was built for structure-on-read processing, making it readily accessible to the typical discovery analytical work done by data scientists. Vendors on the data lake side are working to make complex analytical processing more readily accessible and performant on their platforms.

From the data warehouse side of the race, mature database vendors built for structured data are working to access both semi-structured data and tiered storage. Columnar MPP databases were built for high-speed analytical processing, and some databases have worked to provide notebook and data science access to their platforms. In addition, it is becoming more common to see machine learning algorithms embedded in data warehouse technology for high-performance processing of advanced analytics on structured data.



Hurdles to Overcome

The convergence of the data lake and the data warehouse is not without challenges from both sides of the race. On the data lake side of the race, it is extremely difficult to build an enterprise-ready database. History dictates that it takes at least ten years to harden database technology and make it safe and reliable enough for enterprise customers. Many data lake vendors currently tout the release of some enterprise database features, but most are a long way off from the deployment of full enterprise-strength database functionality.

On the data warehouse side, it is challenging to break out of the structure built into a database and to open

the technology up for semi-structured data and random discovery of analytical workloads. In addition, it is difficult to loosen the tight storage to compute coupling to work with tiered storage. Leaders among the analytical database vendors are already finding ways to overcome these challenges. EMA believes they will be the winners in the race because of the amount of time it takes for data lake vendors to build a proven and reliable database.

REQUIREMENTS FOR THE UNIFIED ANALYTICS WAREHOUSE

To assess the likely winners in the race for the unified analytics warehouse, it is important to understand the various requirements of modern analytics programs and the unified analytics warehouse.

Data Requirements

The modern enterprise produces and utilizes a broad range of data types. For example, a customer-facing application may capture semi-structured web stream data, mobile app data, raw text from email, location data, and structured data from both marketing and sales automation systems. Customer engagement requires a single platform to easily capture, combine, and analyze these different streams and sets of data. The support of multi-structured data is vital to the success of the unified analytics warehouse.

Today's digital enterprise also needs to respond intelligently to real-time business and customer events. Therefore, the UAW must enable organizations to ask questions of data while it is in motion and while it is at rest, or to enable queries across both states of data.

Enterprise Requirements

Both longstanding companies and digital-first companies require specific enterprise capabilities to meet corporate, regulatory, and competitive requirements. These include, but are not limited to, management, orchestration, security, privacy, performance, scalability, agility, resiliency, and affordability. A successful unified analytics warehouse must be enterprise-ready.

Infrastructure Requirements

Historically, businesses moved their data from databases to file systems to save money. Now, they are moving from file systems to object storage. In the world of analytics, it is important to remember that cheap storage is limited. If it is not accessible for analysis, cheap is not enough. For this reason, the unified analytics warehouse must be able to provide a rich and consistent set of analytical capabilities across all storage tiers. More advanced UAWs will automate the movement of data in and out of file systems and object storage when needed.

Cloud Requirements

Recent EMA research showed that 53% of all data is now in the cloud. Because of this massive shift of data, both multi-cloud and hybrid support is essential for the unified analytics warehouse. A legitimate UAW will enable management of these different systems in a single pane of glass for all environments. Additionally, in the best-case scenario, the unified analytics warehouse software should live and act in the same way across different cloud providers and on-premises.

Analytical Processing Requirements

Analytics are the primary use case of the unified analytics warehouse. Therefore, at a minimum, the unified analytics warehouse must include a common set of prebuilt analytical algorithms and functions to address every step of the machine learning process. Those algorithms should be embedded in the platform for ease of use and high performance on the largest data volumes, without any downsampling of data. Additionally, it must support the rapid development and execution of ad-hoc analytics.

User Expectations

At the highest level, users expect the unified analytics warehouse to provide seamless unification of all interactions with data and analytics. Users do not anticipate having to move to different environments to access data or having to use different interfaces to manage diverse data.

Users also expect a platform that allows them to put aside religious beliefs about how analytics should be done. With the UAW, data engineers, data scientists, and data analysts no longer need to fight about who is right and who is wrong. They have a single environment where they can collaborate for the greater good of the enterprise.

Users expect the unified analytics warehouse to provide seamless unification of all interactions with data and analytics.

To support a unified data workforce, the UAW must also support a broad set of different approaches to analytics. The data scientist must be able to use R, Python, and notebooks to execute discovery analytics or advanced analytics like machine learning on multi-structured data. The platform must enable easy-to-access, high-performance analytics via prebuilt embedded algorithms. It must be simple to combine these analytics for greater insight or ask questions of data in near-real time.

VERTICA – CROSSING THE FINISH LINE

EMA conducted a full product review of the newly released Vertica 10 to determine existing capabilities for the emerging unified analytics warehouse. At the core of the Vertica strategy is the idea that the platform becomes the unifier and maximizer of the underlying infrastructure.

In addition to the following technical and business review, it is important to note that Vertica was founded in 2005 and released their analytic database in 2007, giving their platform 13 years of database hardening and maturity as of 2020.

Analysis of Multi-Structured Data

Vertica 10 introduced expanded support for the analysis of semi-structured data types, especially the complex data types found in Maps, Arrays, and Structs in Parquet data. Instead of replicating the data in Vertica to run a query, Vertica can access the data directly in HDFS or S3 object storage. This eliminates the need for data storage duplication and enables much quicker answers to questions requiring both data stored in Vertica and in other data platforms.

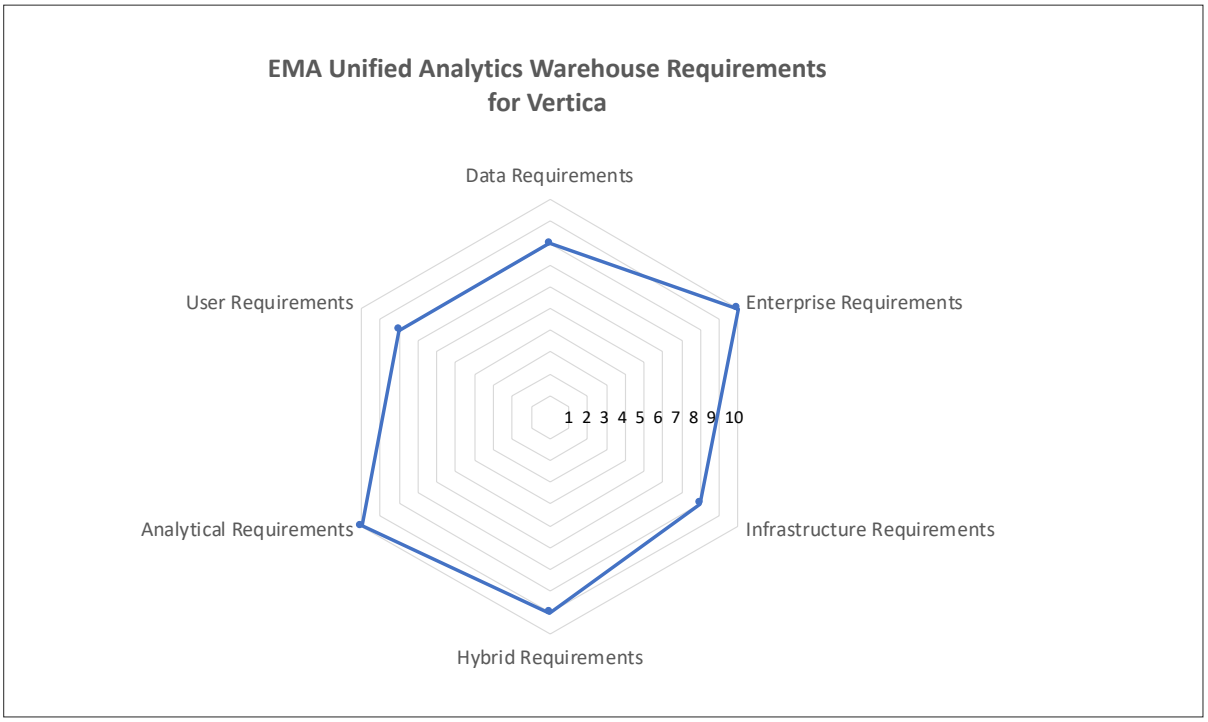
Vertica 10 also supports broad flexibility in how users store data and run queries. Users can query the data in the disparate systems or import it to Vertica's high-performance storage for even faster query response. Since the data can be stored in Parquet on S3 or HDFS, this capability spans the data lake to data warehouse divide with common support for multi-structured data.

Fortunately for Vertica, the company contributed heavily to the Apache Parquet open-source project when it first began. They built the C++ versions for both Parquet and ORC. Their contribution ensured that both formats supported intelligent pushdown of analytical and transformational operations. In addition, Vertica maintains the ability to write out to Parquet, a function generally done only by data lake vendors.

Vertica applies high-performance techniques to complex data types. Instead of exploding them like other platforms, bogging down analyses, they leave the complex data types in their original format. Vertica was built to operate this way. For example, a JSON file does not need to be parsed; it is simply stored in Parquet. Vertica analyzes it natively. Support for multi-structured data is the biggest challenge to achieving a unified analytics warehouse deployment from the data warehouse side, and Vertica has overcome the challenge.

Extensive Enterprise Readiness

With more than ten years of hardening and maturity built into Vertica, there are no questions regarding enterprise readiness, especially within the database platform. Every security and privacy capability has been fully tested. The database is already proven and in use in companies where corporate, regulatory, scalability, and resiliency requirements are paramount. Management, orchestration, and agility objectives are met with singular management across multi-cloud and hybrid environments. Enterprise readiness is the most difficult barrier to entry for data lake vendors wanting to deploy a unified analytics warehouse. This is one of Vertica's greatest strengths.



Full Infrastructure Investment Leverage

By completely separating compute from storage when the customer requires, Vertica provides the best possible solution for leveraging investments in infrastructure. This total separation, combined with the ability to query data in any tier of storage, creates a tremendous advantage in pricing against vendors that provide a combination of their own analytical compute and storage in the cloud.

Data-intensive analytical applications benefit from the use of multi-tiered storage of data. Compute-intensive applications benefit from the high performance of the Vertica analytical database platform. Users can leverage infrastructure investments on-premises, in different clouds, on Apache Hadoop, or as a hybrid model.

One Software for Multi-Cloud and Hybrid

As a software-only vendor, Vertica provides one version of their software platform for all environments. In different environments, Vertica has the exact same set of capabilities. This offers users complete flexibility for on-premises, big data, cloud, and hybrid deployments. Vertica exceeds expectations for multi-cloud and hybrid support by making it seamless to move applications from one environment to another.

In addition, with the ability to license the software once, an enterprise license can be used for all environments, including test, development, and high availability at no additional cost. Vertica supports the deployment of a unified analytics warehouse for almost any scenario, which is especially important when change is inevitable.

Deep Analytical Roots

Vertica has supported the broadest set of analytical functions spanning event and time series, pattern matching, geospatial, and end-to-end in-database machine learning workflows for a long time and continually builds more. Vertica was purpose-built as an analytical platform.

Vertica 10 expanded the platform's ability to support the multifaceted work of data scientists, including the use of artificial intelligence and machine learning. Users can both import and export Predictive Model Markup Language (PMML) models. Operational options include: train models in high-performance Vertica and export them for scoring, train TensorFlow models on GPUs and put them into production on Vertica, train models on any data science framework and operationalize in Vertica, or manage and run models as first-class citizens, just like tables, in the database.

Broad User Support

Vertica 10 provides a single platform to meet a broad set of unified analytics warehouse requirements for data engineers, data analysts, and data scientists working on all types of data. The data scientist can use R, Python, and notebooks to execute discovery analytics or advanced analytics like machine learning on multi-structured data. The data engineer can provision data and analytics for data analysts.

On the same platform, the data analyst validates the data and analytics for accuracy. The business analyst can determine which data and model combination provides the best value creation. Finally, the model can be operationalized in Vertica. This broad user support is the definition of the unified analytics warehouse: easy to access, high-performance analytics for all data, all analytics, and all data professionals.

ABOUT VERTICA

Trusted by AT&T, Cerner, Uber, The Trade Desk, and many other data-driven organizations, Vertica is the unified analytics warehouse that solves three current market challenges that every organization faces. Despite the disappointment in Hadoop, HDFS data lakes represent a very significant investment for many companies, but the value is not equivalent to original expectations. Combined with the explosion of cloud object storage, organizations struggle even more to unify their data. In addition, organizations favor a multi-cloud or hybrid cloud and on-premises deployment strategy as they face the reality of cloud vendor lock-in, costs, and migration challenges. Finally, machine learning is no longer a data science project, but must be put into production to deliver the predictive analytics information in time to allow proactive actions.

Vertica is a Unified Analytics Warehouse that:

- Unifies HDFS data and Object Storage data lakes to capitalize on storage investments and maximize business value.
- Unifies a company's deployment options spanning multi-cloud and on-premises to embrace cloud innovations, prevent lock-in, and meet regulatory and security requirements.
- Unifies the data science community and the business analyst and IT community, enabling each to continue to use their preferred tools and languages while operationalizing machine learning at scale for real-time predictive analytics.

Visit www.vertica.com for more information.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates® (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help EMA's clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals, and IT vendors at www.enterprisemanagement.com or blog.enterprisemanagement.com. You can also follow EMA on Twitter, Facebook, or LinkedIn.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
www.enterprisemanagement.com
3996.06112020