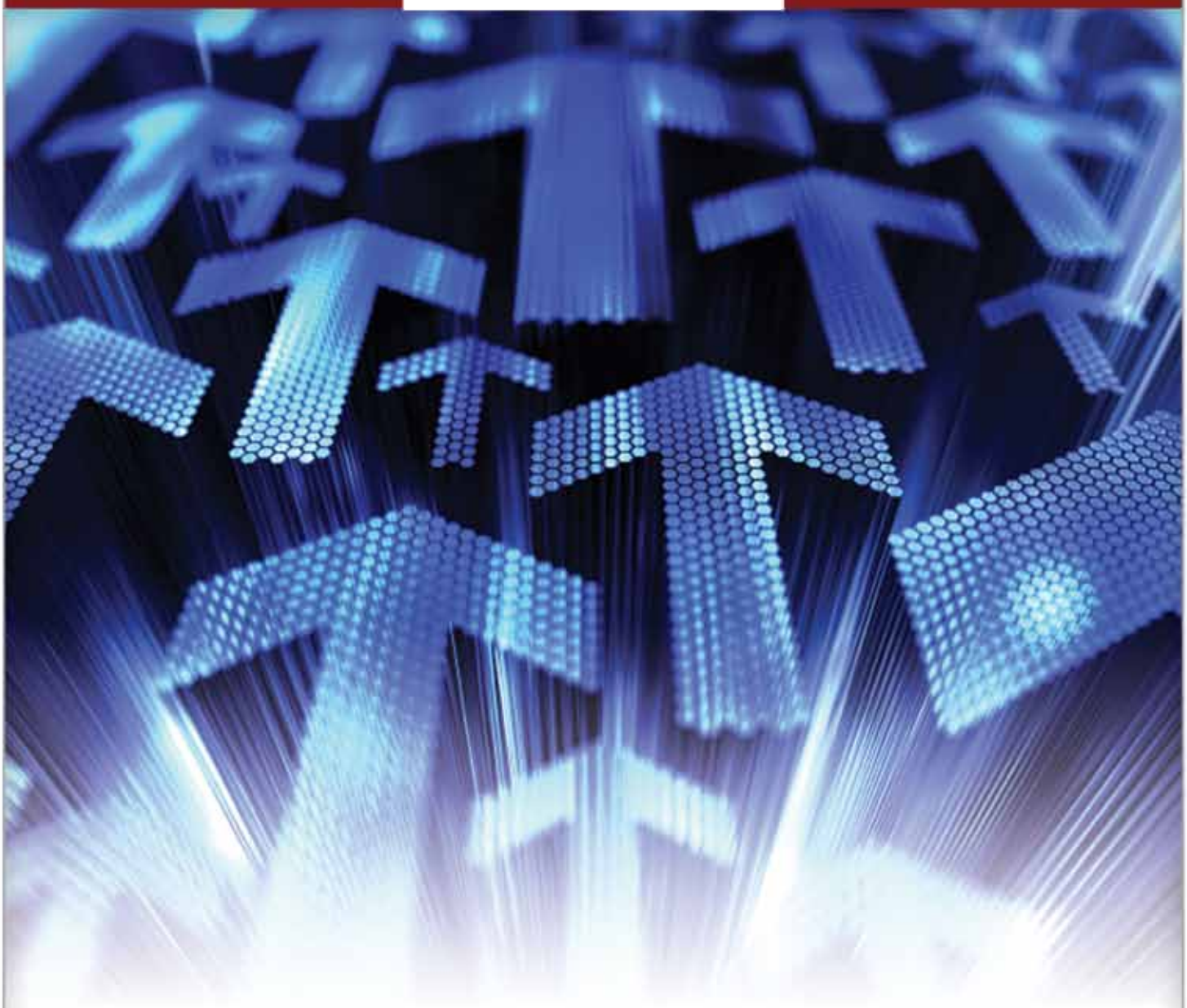
DataDirect™ NETWORKS FEDERAL



DDN | Solution Brief

VERTICA An HP Company

Overcoming >

The Big Data Technology Hurdle

Turning Data into Answers with DDN & Vertica

ddn.com

INFORMATION IN MOTION™

Executive Summary

DataDirect Networks and Vertica have collaborated to implement a joint solution that provides the fastest analytics solution for "Big Data" challenges on the market. Our solution was recently validated by a mutual public sector client with extremely large data requirements. The client recognized the value of the solution based on its ability to provide extremely fast time to actionable information from large datasets, compiled from hundreds or thousands of sensors, social media channels and other data collection sources.

The validation testing was conducted on a test database of over **one trillion rows of data**.

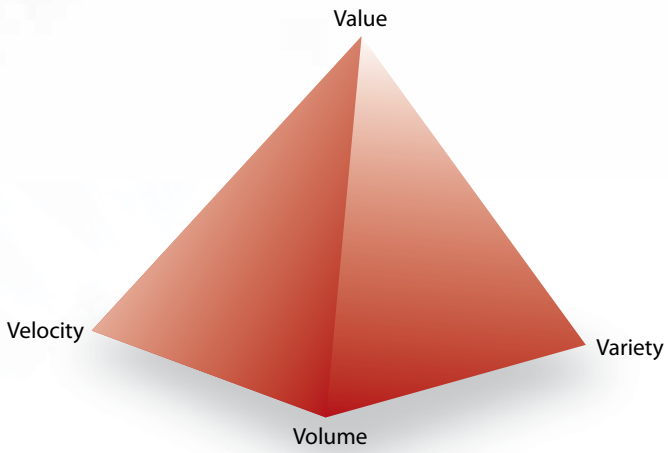
The **DDN/Vertica** joint solution exceeded their requirements based on:

- Data ingest speed
- Solution scalability
- Query speed
- System flexibility & ease of management

Simple queries were completed on the Trillion row database in as little as two seconds, with equally impressive results on more complex queries.

The following document provides information on current big data challenges; the strengths of the joint **DDN/Vertica** solution and how it will be able to assist organizations extract better information faster – from extremely large data sets.

> *"The amount of data in our world has been exploding. Companies capture trillions of bytes of information about their customers, suppliers and operations, and millions of networked sensors are being embedded in the physical world in devices, such as mobile phones and automobiles, sensing, creating and communicating data. Multimedia and individuals with smartphones and on social networking sites will continue to fuel exponential growth. Big data – large pools of data that can be captured, communicated, aggregated, stored, and analyzed – is now part of every sector and function of the global economy."*
>
> **– McKinsey Global Institute**
> Big data: The next frontier for innovation, competition, and productivity



In the era of big data, analytical analysis presents amazing new opportunities. With the volume, velocity and variety of data created by a multitude of sensors and human generated sources, analysts have the ability to accumulate quantities of data that would have been unimaginable just a few years ago. In addition to new data collection capabilities, analysts also have a new generation of analytical tools and human expertise that is rapidly being developed to extract value from these new data sources. The results of these advances are providing today's analysts with unprecedented ability to make real-time data-driven decisions unlike any time before.

When building a system for real-time decision making, it is critical to optimize each element of the system architecture for balanced performance and long-term scalability. Today's data velocity and variety can challenge conventional system technology, and unpredictable volume requirements add new complexities as customers consider how they will respond to tomorrow's data generators. An imbalanced analytics system can result in bottlenecks that reduce the value of analytics tools and the enterprise data warehouses (EDW) that power them. Most critically, balanced infrastructure is key to ensuring the extraction of value to organizations, at the time they need it, and within the budget they can afford. As such, it's critical for organizations to ensure that at either end of the ingest/egest process, that they build scalable storage and storage processing technology to complement today's high-speed data warehousing and data analytics technology.

Big Data Analytics

It is increasingly commonplace for today's data-driven enterprise to be managing data volumes that scale into the petabyte range. Facebook, for example, shares more than 30 billion pieces of information a month. As organizations look to extract value from these large volumes, they must process both structured and unstructured content to develop a comprehensive understanding of both data (structured information) and sentiment (unstructured

data, or file-based data). As conventional data warehousing tools have proven to be challenged by the volume of information being processed, attempts have been made to minimize the impact of these challenges by either reducing the data being processed (summary data analysis), or increasing the speed of analysis.

Initially brought together by a mutual customer – DataDirect Networks (DDN) and Vertica, An HP Company, have now embarked on a mission to liberate EDW users from the bottlenecks associated with conventional data warehousing and data analytics. We introduce a highly-scalable reference platform, designed to scale with the requirements of today's and tomorrow's Big Data value extraction. The joint solution can scale to ingest terabytes of data a second from 1000s of different sensors or other data sources. It can process this data up to 60% faster than standard analytics tools and manage 10's of trillions of rows of data and 100's of billions of inserts, daily.

DDN + Vertica: Real Life Results for Massively Scalable Big Data

The Requirement: The Chief Technology Office of a mutual public sector client came to Vertica and DDN to solve their Big Data Technology hurdle – "how to run real-time analytics against an impossibly large and diverse data set on the agency's Cloud infrastructure?"

Specifically, the client needed to support a cloud-based analytics cluster using commodity-based x86 Linux servers and standard networking hardware to analyze data from hundreds of sources, including social media. The system needed to upload structured and unstructured data into a "Cloud" data layer and support real-time ingest and queries of the data. The test environment needed to ingest and extract entities from 42,000 documents an hour and maintain query responses of less than 1 second. The system also needed to linearly scale to support the inserting of 10's of millions of documents an hour, while providing real-time query access to the database while ingesting.

Critical Elements

The client recognized there were four fundamental capabilities that needed to be brought together in a single solution to meet their requirements:

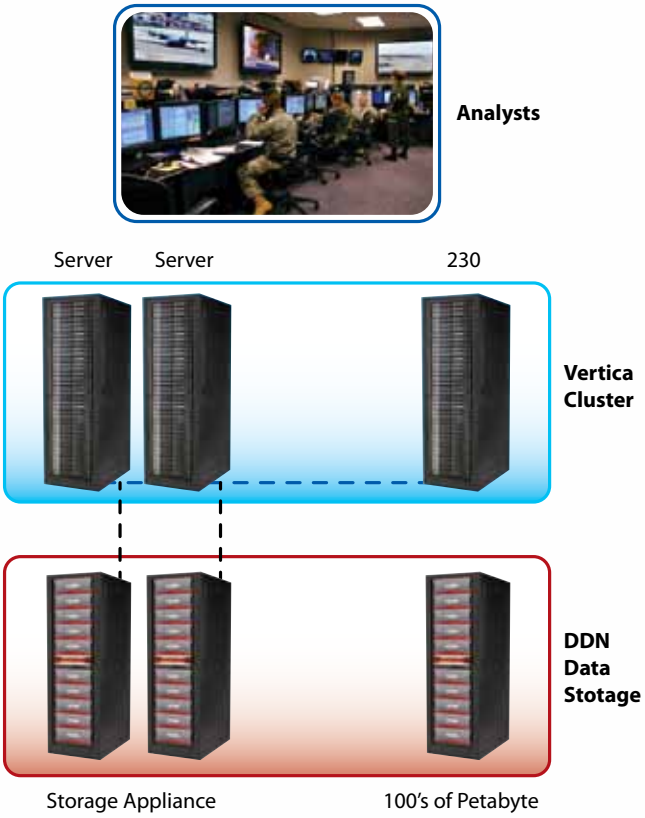
- Ingest speed
- Scalability
- Query speed
- System flexibility & ease of management

Ingest speeds – Having deployed dozens of systems over 300GB/s (with some scaling over a TB/s), DDN is the unrivaled leader in data ingest. Vertica is an ideal complement to these throughput speeds. Vertica's parallel loading capabilities enable the database to scale to accommodate ingest speeds well in excess of a TB/s. Utilizing the DDN Storage Fusion Architecture™ (SFA) systems, in conjunction with Vertica, the client was able to load information into

their Cloud-based data layer nonstop – while simultaneously allowing access to the data for rich, real-time queries. As the critical first step in the joint solution, the DDN platform is able to serve data to the application at speeds that maximize the overall performance of the Vertica solution.

Scalability – Both DDN and Vertica are recognized for their ability to scale beyond traditional solutions in their respective fields. The **DDN/Vertica** joint solution can grow to hundreds of nodes supporting multi-petabyte environments. More importantly, the system can scale modularly. As the customer adds additional database content or users, both DDN and Vertica can add additional nodes in line with their requirements without taking down the system.

Query speeds – Vertica was designed with a unique, time-travel transactional model, that ensures extremely high query concurrency, while simultaneously loading new data into the system. Vertica is often able to load data up to 10x faster than traditional row-stores as a result of this design. Utilizing the DDN platform, the Vertica solution can deliver sustained query responses of less than 1 second from the system, even as new data is added.



At maximum capacity the **DDN/Vertica** Solution can scale up to support ingest speeds between 1.4 and 1.8 TB/S*

System flexibility and ease of management – Both DDN and Vertica are able to scale their solutions up or down based on policy, adding or removing nodes in real time. Additionally, DDN is able to scale performance and capacity, independently of each other, for greater flexibility and efficiency.

Test Configuration

DDN | SFA10K-X

- InfiniBand Native Attach ~12GB/s
- 40 64GB SSD Drives
- 560 300GB 15k RPM SAS Drives
- 125.16TB usable to the filing system

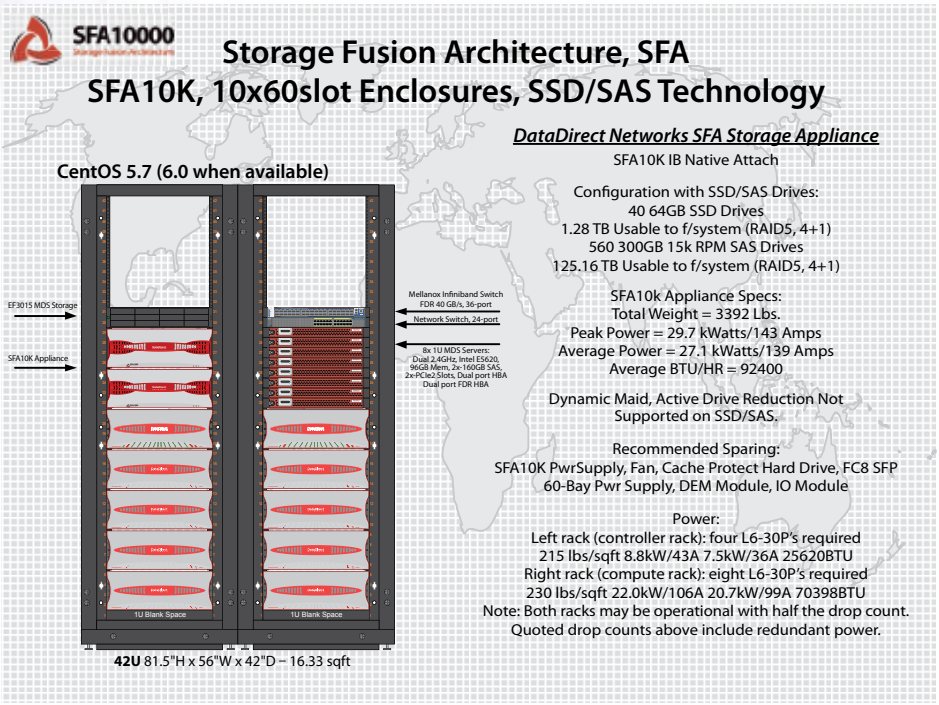
8-Nodes (servers)

- Dual Hexcore per node – Intel Xeon Processor E5620 (12M Cache, 2.40 GHz)
- 96 GB Memory per node
- 40 64GB SSD Drives used for Vertica Catalog Files
- 560 300GB 15K RPM SAS Drives used for Vertica Data Files (RAID 5, 4+1)

Standard GigE Network

- Data transfer and inter-nodal communication over TCP/IP

With this configuration, the solution achieved data load rates of **10-billion rows per hour** or **2.78 million rows per second** (Scales With Additional Servers). Every 10-billion rows of data equated to 900GB of data loaded per hour or 250MB of data per second. The customer was able to load, query and modify the database schema in parallel, while in Full production.



Results

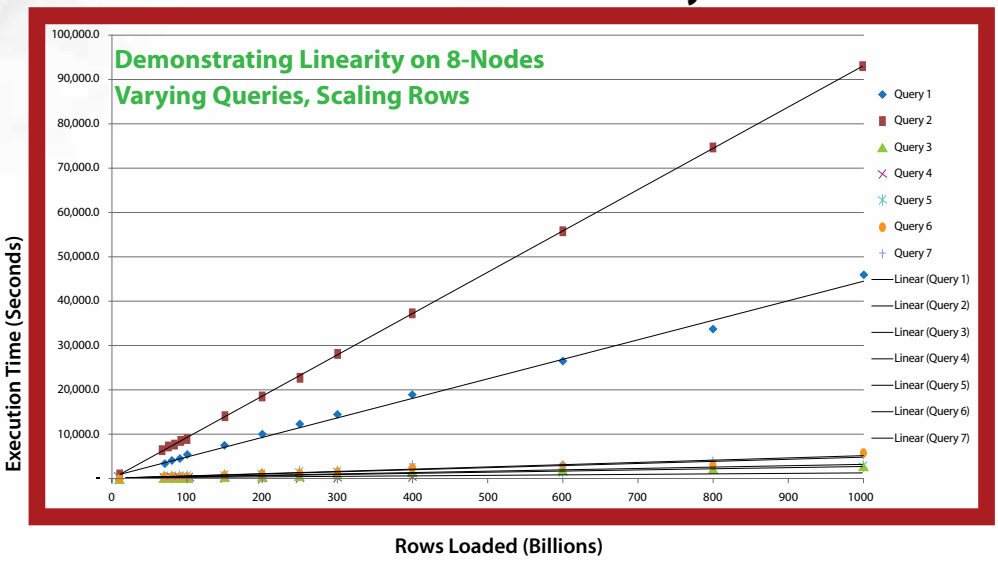
The trial was performed on a test database consisting of **one trillion** rows of data. To demonstrate integration with a variety of querying, reporting, and visualization tools, including: DBVisualizer (an open source database access product), Microsoft Excel, and Tableau Desktop.

The benchmark performed a series of queries over an increasing data volume in the database. These queries reflected a varied SQL workload, from data aggregation to data filtering – to filtering on the data aggregation. The results proved the joint solution's query speed scales linearly (see chart below). Linearity of scale is also achieved by adding additional servers to the cluster for additional performance.

Simple queries against the one trillion row database were completed between **2 and 20 seconds**.

Complex queries ranged in response times from 20 minutes to 20 hours, depending upon the volume of data that was aggregated.

Test Performance – Solution Linearity on 8-Nodes



The Vertica Advantage

The Vertica Analytics Platform is the only database built from scratch to handle today's heavy business intelligence workloads. In customer benchmarks, Vertica has been shown to manage petabytes of data and answers queries 50x to 1000x times faster than competing row-oriented databases and specialized analytics hardware.

During the last 30 years, there has been little innovation in database management systems (DBMS) to keep pace with the growing volume of data produced. Performing ad hoc queries on such multi-terabyte or petabyte databases does not come naturally for existing DBMS systems, which use a row-oriented design optimized for write-intensive transaction processing workloads rather than for read-intensive analytics workloads. Desperate for better performance, row-oriented DBMS customers spend millions of dollars annually on stop-gap measures like adding database administrator resources, creating and maintaining OLAP cubes, or replacing their DBMS with expensive, proprietary data warehouse hardware.

Vertica developed a break-through SQL database that offers all sizes of companies a competitive advantage and cost-effective ways to analyze terabytes or petabytes of information. Here is what sets Vertica apart:

- Radically improved database performance
- Painless scalability
- Simplified database administration – administration tools that empower and enable

Vertica's SQL-based platform provides blazingly fast query performance for databases scaling from hundreds of gigabytes to multiple petabytes, running on industry-standard hardware, virtual machines (private cloud, Ghost Machine) or on a public cloud. What makes the Vertica Analytics Platform architecture so unique is that it combines the database industry's most significant recent innovations into a single database for the first time:

- Columnar database architecture
- Distributed, grid computing, and shared-nothing architecture
- Aggressive data compression
- Automatic physical database design
- Automatic "log-less" recovery by query
- Scale-out MPP architecture on industry-standard servers, virtual machines, or on the cloud
- Hybrid transaction architecture that supports concurrent, parallelized querying and loading of data
- Multiple physical sort orders and joins of related datasets ("projections")
- Native High Availability without hardware redundancy
- Universal connectivity via JDBC/ODBC and SQL to BI applications, ETL, and querying/reporting tools

The DataDirect Networks (DDN) Advantage

The DataDirect Networks Storage Fusion Architecture (SFA) combines unprecedented IOPS and bandwidth performance with highly efficient capacity management to maximize application performance and minimize overall total cost of storage system ownership for data-intensive environments. The scalability and performance of DDN storage systems, including the fastest storage systems in the world with aggregate performance beyond 1 Terabyte per Second, helps organizations maximize the value of information and accelerates analysis and critical decision making.

The DDN solution supports storage arrays, file systems and object storage appliances for the cloud to the world's most data-intensive environments. These scalable, highly efficient storage solutions enable our customers to accelerate time to results, scale simply as data sets continue to grow and gain competitive advantage through resolving performance and capacity scaling challenges. By optimizing each element of the I/O environment for performance, capacity and data center efficiency – DDN solutions deliver the highest levels of performance and fastest time to results. The DDN difference:

- 800% Faster than competitive solutions – Single systems provide up to 40GB/s and reaching 1TB/s can be achieved in just 25 systems
- Enormous Internal Bandwidth – The DDN architecture protects performance during drive rebuilds, while competing systems lose as much as 40%
- In-Storage Processing™ – enables latency-sensitive applications to live right inside the storage appliance

DDN | SFA is designed to derive peak performance from data intensive applications. With a massive I/O infrastructure and multi disk technologies that maximize system performance and lower storage investment costs, the DDN architecture delivers unrivaled benefits:

- Supports multiple high speed connectivity protocols, such as InfiniBand, Fibre Channel and Ethernet to enable the highest data ingest rates in the industry
- Balanced, Best-In-Class Storage Performance – Maximum performance and efficiency for both highly transactional and high-bandwidth applications
- ReACT™ Intelligent Cache Management – Optimizes writes in real-time: sequential data goes directly to disk media while small, random IO utilizes extremely fast cache
- Read I/O Quality of Service – Read IO doesn't suffer due to a single, unresponsive disk
- DirectProtect™ Real-Time Error Detection & Correction – Increases data resiliency and reliability with little performance impact
- Storage Fusion Fabric™ Unparalleled Back-end SSD Support – Fully utilizes SSDs for unprecedented levels of sustained random IOPS
- Journaled Drive Rebuilds – Reduces rebuild times by only requiring new/changed blocks to be written to recoverable drives

Combining Strength with Strength

Combining the industry's leading big data analytics database, with the leading data storage solution for data intensive computing environments, DDN and Vertica have delivered unmatched levels of data analytics capabilities. The joint solution can scale to ingest over a terabyte of data a second from multiple, disparate data sources, including sensors. It can process this data up to 60% faster than standard analytics tools, and manage 10s of trillions of rows of data with 100s of billions of inserts a day.

In short, this partnership combines the industry's fastest analytics platform with the industry's fastest storage engine. The resulting solution can manage unprecedented amounts of data, scaling up to multiple petabytes, and process this data at unmatched speeds. More importantly, when you combine Vertica's ability to lower disk space requirements with DDN's ability to scale performance and capacity independently, you can deliver optimal performance with far greater efficiency.

Summary

Tasked with a unique combination of big data challenges, this Agency client understands the critical difference between data and information. By combining the biggest database with the fastest storage solution, this client is able to convert enormous amounts of data into actionable intelligence to solve some of the most challenging data problems in the world.

While this client is clearly an early adopter and a high-end user of data and analysis, their requirements to extract information from large data sets are by no means unique. The joint **DDN/Vertica** solution delivers unmatched performance and decision making capabilities for any organization trying to extract vital information from huge, ever-changing, data sets derived from a wide variety of data sources. With the ability to scale in-line with requirements up to multi petabyte environments, the **DDN/Vertica** solution can process trillions of rows of data from hundreds or thousands of data sources. Utilizing a solution that can deliver data to the analytics, (or better yet bring the application to the data through in-storage processing), process it in real time, and scale to virtually any size can mean all the difference between being able to store data and being able to turn that data into actionable information.

DataDirect NETWORKS FEDERAL

DDN FEDERAL | About DataDirect Networks

DataDirect Networks (DDN) is the world leader in massively scalable storage. We are the leading provider of data storage and processing solutions and professional services that enable content-rich and high growth IT environments to achieve the highest levels of systems scalability, efficiency and simplicity. DDN enables enterprises to extract value and deliver results from their information. Our customers include the world's leading online content and social networking providers, high performance cloud and grid computing, life sciences, media production organizations and security & intelligence organizations.

Deployed in thousands of mission critical environments worldwide, DDN's solutions have been designed, engineered and proven in the world's most scalable data centers to ensure competitive business advantage for today's information powered enterprise.

For more information, go to www.ddn.com or call +1.800.837.2298

ddn.com

INFORMATION IN MOTION™

DataDirect Networks Federal, LLC. | 8320 Guilford Rd, Columbia, MD 21046 | government@ddn.com | +1.800.837.2298