Data Analytics – The New Growth Opportunity for Software Developers

How the Vertica Analytic Database is powering the new wave of commercial software, SaaS and appliance-based applications and creating new value and competitive differentiation for solution developers and their customers

Overview

Regulatory compliance, increased competition, and other pressures have created an insatiable need for companies to accumulate and analyze large, fast-growing quantities of data such as:

- Telecommunications call detail records (CDRs)
- IT/Network event history
- Financial trade (quote and tick) history
- Web logs & click streams for marketing and fraud analytics
- Compliance and other historical logs

"Better than Free"

said nMetrics (www.nmetrics.com.au) CTO, Dave Britt, of Vertica after replacing open-source PostgreSQL with Vertica inside their network performance monitoring appliance.

Vertica was able to store and analyze 1 week of network traffic data faster and on less hardware than PostgreSQL could manage 1 hour's worth.

This presents a major market opportunity for enterprise software vendors and software as a service (SaaS) companies. They can profit by creating analytic data management features or entirely new applications that put customers on a faster path to better data-driven decision making. Offering such BI capabilities enables application vendors to not only keep a larger share of their customer's budget, but also greatly improves customer satisfaction.

To offer such capabilities though, many solution providers are performing unnatural acts with row-oriented databases that were never meant for heavy analytic workloads, or alternatively they are forced to build their own data management systems. Both of these approaches are costly and outside the core competency of most enterprise software companies. Fortunately for them, there is now an alternative that is simple, cost effective, and offers dramatically better performance.

The column-oriented Vertica Analytic Database is a relational SQL database built specifically to handle today's analytic workloads. Unlike commercial and open-source row-stores, which were designed decades ago to support OLTP workloads, Vertica provides ISVs and their customers with:

- Superior query and analytic performance
- Better compression and more efficient data storage/resource management
- Flexibility and scalability to store and query more raw data ad hoc
- Better load throughput and concurrency with querying
- Less DBA overhead and tuning

This paper compares Vertica with alternative embedded database options available to analytic application developers, such as:

- Commercial OLTP databases
- Open source OLTP databases
- Proprietary data warehouse appliance hardware
- Proprietary ISV-built databases

The New Growth Opportunity for Software Vendors: Data Analytics

Twenty years ago, the business transaction processing software market was in transition. Instead of building and maintaining ERP, CRM and other applications internally using large, expensive IT teams, companies began purchasing commercial off-the-shelf (COTS) applications from new vendors like SAP, JD Edwards and Peoplesoft. A multi-billion dollar software business was born.

Ten years later, in the early 2000s, the trend was further transformed by Salesforce.com and others who innovated the Software as a Service (SaaS) model and enabled companies to implement new applications without traditional data center costs or time-consuming cap-ex budget approval processes.

The major market opportunity today for enterprise software vendors and SaaS companies is in the analytic application arena. The funding for these initiatives is available (as evidenced by the billions spent annually on custom data warehousing technology and services), and software companies can profit by putting customers on a faster path to better data-driven decision making. Such solutions enable customers to gain insights for a competitive edge, to reduce risk exposure, and/ or increase profitability.

Enterprise software and SaaS vendors have benefited greatly from underlying database management systems allowing architects and developers to focus on their core competencies at the application level without worrying about the underlying data management. Historically, these traditional OLTP row-oriented engines could handle the transactional data entry nature of most applications. However, today's analytic applications must continuously and simultaneously load and query against massive volumes of information.

Analytic Applications Require an Analytic Database Management System (DBMS)

As independent software vendors (ISVs) shift new product development and innovation from data collection to operational analytics and business intelligence (BI), they face customer requirements that are much different than the OLTP workloads of the past.

The grid-based, column-oriented Vertica Analytic Database is the only DBMS built specifically to handle analytic workloads, and it provides ISVs and their customers with highly differentiated capabilities, performance and economics.

Analytic Data Management Requirements	
Focus	Querying, not Updating
Data volume	Large, fast growing—users want more detail and more history
Access	Ad-hoc analysis of real-time data, not yesterday's batch
Performance	Answers in seconds, not hours
User load	Many people (esp. for SaaS), not just a few power users
IT Cost	Less hardware and manual tuning overhead the better

Comparing Embedded DBMS Alternatives for Analytic Applications

The table on the following page provides a high-level comparison of different embedded database alternatives for analytic applications. ● = *Excellent* ⊙ = *Fair/Variable* ○ = *Poor*

	Analytic DBMS	OLTP DBMS	Proprietary data warehouse hardware	Custom-built database
ARCHITECTURE				
Intended workload	Analytic	OLTP	Analytic	Analytic
Data storage	Column-oriented	Row-oriented	Row-oriented	Custom
Hardware	Shared-nothing grid	Large SMP servers, SAN	Proprietary	Variable
Sample products	Vertica Analytic DBMS	Oracle, MySQL SQLServer PostgreSQL	Netezza Oracle/Exadata	Based on file system, etc.
PERFORMANCE & EFFICIENCY				
Query speed	●	○	⊙	⊙
Loading (Insert) speed	●	⊙	●	⊙
UPDATE speed	○	●	⊙	⊙
Data Compression	●	○	○	⊙
REPORTING FUNCTIONALITY				
Ad-hoc analysis	●	○	●	⊙
Real-time analysis	●	○	●	⊙
BI tool connectivity	●	●	●	○
SCALABILITY				
Large data volumes	●	○	●	⊙
Large user volumes	●	○	●	⊙
Ease of scalability	●	○	⊙	⊙
COSTS				
Admin overhead	●	○	●	●
Hardware costs	●	○	○	⊙
Time to market	●	⊙	●	○
Replication/HA costs	●	○	⊙	○
ISV BUSINESS MODEL SUPPORT				
Deployment flexibility	●	⊙	○	●
Licensing flexibility	●	●	○	●

Architecture – Columns, Compression and Clustering

Because it's the only DBMS built specifically for supporting analytic workloads, Vertica is the most highly differentiated of the commercial alternatives described in this paper. Vertica differs from OLTP DBMS and proprietary appliances (which typically embed row-store DBMSs) by grouping data together on disk by column rather than by row (i.e., so that the next piece of data read off disk is the next attribute in a column, not the next attribute in a row). This enables Vertica to read just the columns referenced by the query, instead of having to scan the whole table as row-oriented databases must do. This speeds up query processing dramatically by reducing disk I/O.

By grouping data together on disk by column, Vertica creates the perfect scenario for data compression—lots of similar or repetitive values can be compressed very aggressively. Vertica features a library of many compression algorithms, which it applies automatically based on data type. Typically, a Vertica database occupies up to 90% less disk space than the data loaded into it. This not only lowers storage costs, but also speeds up querying by further reducing disk I/O.

The last architectural point has to do with hardware architecture. Vertica is a "shared-nothing," distributed database designed to work on clusters of inexpensive, off-the-shelf servers, and its performance is scaled simply by adding new servers to the cluster. It is also worth noting that Vertica can be embedded and run on a single node appliance as well. Vertica's grid architecture reduces hardware and scaling costs substantially (by 70%-90%) when you compare it to traditional databases that require "big iron" servers with many CPUs; and storage area networks (SANs). Clustering also speeds up performance by parallelizing querying and loading across the nodes in the cluster for higher throughput.

These 3 architectural differences—column orientation, compression and the ability to run and scale out on shared-nothing clusters of inexpensive servers are what fundamentally enable Vertica-based analytic applications to:

- Meet more aggressive performance service level agreements (SLAs) than the competition
- Deliver highly differentiated analytic capabilities to customers
- Scale seamlessly and offer many more users access to much more data
- Cost less to develop and run, thus boosting ISV profitability and/or reducing customer costs.

Meet More Aggressive Performance SLAs with Vertica

Vertica has been specifically designed to support analytic workloads, which primarily comprise querying (often ad-hoc) and the insertion of new data (in bulk or on a constant "trickling" basis for real-time analytics). Because columnar storage and aggressive compression drastically reduce disk I/O, column-oriented databases like Vertica deliver orders of magnitude faster performance than row-oriented databases when handling query-intensive analytic workloads.

To help quantify the advantage for you, the following tables contain performance benchmark results conducted by Vertica ISV/SaaS customers and are based on their real-life data sets:

Vertica manages >25x more data 41x faster than PostgreSQL in a network monitoring appliance

Network Performance Reporting Appliance	Vertica® Analytic Database	PostgreSQL	Vertica Advantage
Benchmark Data	Test 1: 6 hours of network data (425 million SNMP records) Test 2: 6 days of network data (10 billion SNMP records)		
Production Data	1 month of network data (over 30 billion SNMP records)		
Benchmark Hardware	(1) Dell 2850 Dual Xeon, 4GB RAM, 1TB local disk		Manages 25x more data on the same hardware
Test 1–6 hours of data (mean query time)	1.08 seconds	303.77 seconds	**304x faster**
Test2–6 days of data (mean query time)	7.47 seconds	Too lengthy to run	Queries 6 *days* of data 41x faster than PostgreSQL queries 6 hours' worth

Vertica manages click stream data 33x faster on hardware costing $900K less at a SaaS provider

Web Marketing Analytics SaaS	Vertica® Analytic Database	Proprietary Data Warehouse Hardware	Vertica Advantage
Data Set	25M Rows of Click stream data		
Benchmark Hardware	(4) servers – 2 x dual core x86 CPU, 8GB RAM, 7x146 HDD	Specialized Appliance Hardware - 112 CPU-disk pairs	>$900K hardware savings; no proprietary hardware
Mean Query Time (7 queries)	7 seconds	233 seconds	Vertica is 33x faster

These results (and many others found at www.vertica.com/benchmarks) show that vendors who embed Vertica in their analytic solutions gain a sales and marketing edge by being able to raise prospects' SLA requirements much higher than the competition.

Vertica is also able to load new data at very high speeds as well. In fact, it took a world record[1] breaking, 57 minutes to load 5.4TB of TPC-H data (generated retail sales data) into Vertica running on an HP BladeSystem C7000. Vertica achieves fast loading by caching new, incoming data into a memory-based store, in which the data is stored in columns and can be queried, but is neither compressed nor sorted yet. Then, an asynchronous process called the Tuple Mover, moves batches of data out of memory and into disk-based containers, compressing and sorting the columns in the process. Thus, the performance cost of sorting and compression is shared and amortized across many tuples. Thirdly, since Vertica databases can be partitioned across a grid of shared-nothing servers, Vertica distributes the INSERTs evenly across the cluster based on a partitioning key. If the INSERT rate grows, more nodes can be added to cope with the increase.

Deliver Highly Differentiated Analytic Capabilities to Customers

The performance advantage explained above enables companies that embed the Vertica Analytic Database to really stand out from the competition and deliver higher value to customers. There are several areas where functionality tends to be improved and differentiated with Vertica:

- Ad-hoc access to years of historic data, not just weeks or months
- Analyze data at any level of granularity, not just pre-calculated summary data
- Real-time data loading and analysis

Here are how these capabilities differentiated the following Vertica-based analytic solutions.

Vertica's fast query performance and data compression enabled nMetrics to collect and analyze over 25x more data than the previous PostgreSQL-based version of their network monitoring appliance. As a result, nMetrics customers such as Qantas, Telstra, ABN-Amro and others are now able to economically:

- Find slowly developing network problems (that only appear anomalous when studied within the context of days or weeks of data)
- Analyze network activity at lower levels of detail – for example, not just see how much data is passing through port 80, but know how much of it is Skype, video, web traffic, etc. to help customers form better network application usage policies and controls.

Migrating their call detail record (CDR) data warehouse to Vertica enabled the following new capabilities in NetworkIP's web-based, profitability/CDR analysis dashboard used by marketers at prepaid calling services retailers:

- **Ad-hoc data analysis** – at any level of detail (down to individual calls). Previously, users could only access summary data.
- **Real-time analytics** – within minutes of a call, not the day after, which helps marketers change and test calling plan parameters more frequently to maximize profits.
- **More accurate trend analysis** – by having access to more historic data – 2 years of CDR history, not just 90 days, to better compare demand and profitability changes over time.

[1] Visit http://www.ETLWorldRecord.com for more information about the benchmark

License More Users to Concurrently Access More Data with Vertica

Because row-oriented databases were designed to handle update-intensive OLTP and not query-intensive analytic workloads, they often limit the ability of analytic software to provide access to lots of data or to support many concurrent users. The problem is especially acute for SaaS solutions, which must return query results to users in "web time," often within 5 to 10 seconds. These databases (or the analytic applications that embed them) can't scale without being migrated to more expensive, high-end servers and SANs or without costly database administration and tuning overhead, which erodes solution profitability.

Columns, compression and clustering give Vertica the ability to answer more questions faster, store more data and scale "out" over time to handle more users or serve more data simply by adding more servers to the Vertica grid. As a result, Vertica-based analytic applications such as those described below are able to support orders of magnitude more data and more users:

Diio (www.diio.net) hosts a SaaS that aggregates air travel passenger ticket sales data. Users within airline companies (Diio's customer base carries 50% of the world's passenger traffic) and airports use the data to help increase profitability by making better informed decisions about route scheduling and fares, etc. They replaced Oracle with Vertica in order to serve more data to more concurrent users. Here's how the DBMS change improved the customer experience along with Diio's ability to grow their user base faster with lower data center costs.

- Average query time shrank from 20 minutes to 6 seconds under moderately heavy user loads (60 concurrent users).
- Users gained the ability to perform ad-hoc analysis of 1.5 years of historic data instead of 90 days of pre-aggregated information.

VMS (www.vmsinfo.com) hosts a SaaS that aggregates TV advertising data for automotive, telecom, retail fast food and pharmaceutical industries. Delivering this information to marketers through a web application is revolutionary and allows customers to respond to changes in competitors' advertising messages and placement immediately (versus relying on weekly or monthly reports from other ad metric reporting companies). They replaced Oracle with Vertica to enable them to grow their database and user base more rapidly and at lower cost:

- Vertica was shown to process over 30 times as many requests as Oracle…
  - Vertica: 95% of queries ran sub-second and the longest query took 1.9 sec.
  - Oracle: 0% ran sub-second, 55% took over 5 seconds and the longest took 20sec.
- Vertica and the hardware to run it cost 65% less than Oracle license fees alone.

Lower Costs and Increase Profitability with Vertica

Another benefit of building applications on Vertica is the cost savings. Here's how Vertica affects cost compared with the other data management options included in the chart on page 4:

Vertica versus…

- **Commercial OLTP databases** (like Oracle and SQLServer) – As mentioned in several of the benchmarks in this paper, Vertica provides much higher performance than OLTP databases on less expensive hardware. For example, a 5 to 10TB Vertica database would run on a grid of 5 HP ProLiant DL 380 servers (or something comparable) costing approximately $50,000 in total. To support 5 to 10TB, an OLTP database would typically be deployed on an SMP server and a SAN costing between $500K and $1M USD. In addition, since Vertica provides superior performance out of the box and is also self tuning and automates database design for speed and fault tolerance, the number of DBAs required to manage Vertica is typically reduced from 2 to 4 full-time DBAs with Oracle to 1 DBA for Vertica. For a SaaS provider, these savings (along with the data center space and utilities conserved) directly impact the bottom line. For COTS vendors, it becomes a cost benefit which can be passed along to customers to help apply pricing pressure on competitors.
- **Open-source OLTP database** (PostgreSQL, MySQL, et al) – Open source databases are very attractive to ISVs because they don't charge licensing fees (although they will sometimes charge ISVs redistribution license fees). Unfortunately for some ISVs, the savings are short-lived because using open source OLTP row stores results in the high hardware and DBA costs described above for commercial OLTP databases.
- **Proprietary appliance hardware** – Data warehouse appliances such as Netezza often combine custom hardware with open source OLTP databases to deliver "plug-and-play" (i.e., little-to-no DBA overhead required) platform that is faster and less costly than commercial OLTP databases. The drawback is that these appliances are very expensive due to their proprietary nature, and they can also be outgrown, which results in expensive data migration work. Furthermore, it is very difficult to "embed" a standalone hardware device into any application. Vertica's flexibility, small footprint, and ability to scale out over time on shared-nothing grids of inexpensive off-the-shelf servers provides a most more cost-effective platform than the appliances.
- **Custom-built databases** – ISVs are sometimes inclined to build specialized data management solutions for their products in order to a) escape database license fees and/or to b) meet specialized performance requirements that cannot be met by commercial products. If Vertica can meet the performance requirements of the ISV solution, it can result in faster time to market and substantial development savings over time by eliminating the time and cost of building and maintaining a custom DBMS code base. Just as Vertica's core competency is in the DBMS, most application developers would rather not be in the business of designing, developing, and supporting a proprietary data store, especially given how important standard such as SQL have become.

Developing with Vertica

Getting started with the Vertica Analytic Database is easy. It supports standard SQL and can be accessed by commercial DBMS reporting and development tools via JDBC, ODBC, ADO.NET and specific language bindings for Perl, Python and others.

Migrating existing application code to Vertica is also straightforward. Vertica implements the PostgreSQL SQL dialect, so porting applications from PostgreSQL to Vertica tends to go especially smoothly. Vertica also includes some Oracle specific SQL dialect such as DECODE to make it easy to port Oracle applications too.

Applications that are UPDATE intensive (i.e., traditional OLTP workloads) may not be well-suited for Vertica, because Vertica is specially optimized for query-intensive workloads. As a rule of thumb, if more than 10% of the workload is UPDATE oriented (versus query or INSERT commands), then Vertica may not be suitable for the application. Lastly, Vertica does not support triggers, but does support external procedures.

Other than those caveats, developing with Vertica is easy to build into enterprise software, SaaS and appliance-based applications:

- Supports DBMS development standards – JDBC, ODBC, SQL & popular language bindings
- Installation and administration APIs are available for more seamless embedding
- Physical database design, replication and failover are automated to save development time and improve performance and high-availability
- Features well-developed security, multi-tenancy, ACID transactions and time travel (snapshot isolation) functionality to protect customer data
- On-line developer support and resources are available at the Vertica V-Zone

Licensing and Deploying Vertica-based Solutions

Commercial software applications are delivered to customers in a variety of ways, and Vertica offers the deployment and licensing flexibility required to be built into:

- Installed software solutions
- Appliances
- In-house hosted solutions
- Cloud-hosted solutions

Vertica runs on low-cost, industry-standard Linux servers[2]. Its fast performance, shared-nothing architecture and compression enable it to run on much less expensive hardware than any other database, which lowers solution development and hosting costs. In addition, an on-demand version of Vertica hosted in the Amazon Elastic Compute Cloud (Amazon EC2) can be licensed on a monthly basis for any length of time, which allows solution providers to bring new offerings to market faster and less expensively without any in-house data center costs. It also enables new solutions with micro life-spans such as marketing campaign analytics which can be licensed for just the duration of a campaign.

Vertica offers flexible licensing programs to ISVs that are designed to support your licensing model and foster mutual growth and profitability.

[2] Vertica system requirements can be seen here: http://www.vertica.com/_pdf/VerticaSystemRequirements.pdf

Summary

The following table summarizes the benefits Vertica provides developers of commercial analytic software and appliances:

5 Reasons to Consider Embedding the Vertica Analytic Database	
Performance	Guarantee service level agreements (SLAs) competitors cannot match. ▪ Execute queries in under 10 seconds, not 10s of minutes ▪ Support many more concurrent users ▪ Vertica holds the world record for fastest data loading – 5TB in one hour
Competitive Differentiation	Deliver functionality that other databases cannot economically support. ▪ Ad-hoc access to years of historic data, not just weeks or months ▪ Analyze data at any level of granularity, not just pre-calculated summary data ▪ Real-time data loading and analysis
Profitability and Time to Market	Build and run solutions faster and at lower cost by eliminating: ▪ Having to build a specialized data management layer ▪ Time and cost of manual database design and administration ▪ Expensive specialized hardware or “big-iron” servers and SANs ▪ Any in-house data center startup costs at all, by using Vertica for the Cloud
Licensing & Packaging Flexibility	Innovate successful business models with Vertica’s flexible licensing options. ▪ Customized licensing to fit your business model and support your growth ▪ Gives you the freedom to deliver hosted, installed or appliance solutions
Customer Success	Vertica-based applications are widely used within the Global 2000. ▪ At top airline, automotive, telecom, health care, financial and retail companies ▪ Managing call detail records, stock trade history, security & network event data, web click stream logs, consumer data and other large, fast-growing types of information

Additional Resources

If you would like to learn more about the Vertica Analytic Database or if you would like to evaluate it yourself, then visit the following links:

ISV Solution Site	www.vertica.com/embedded	Technical and customer resources for ISVs
Resource Library	www.vertica.com/resourcelibrary	General resources such as white papers, demos, webcasts, system requirements
Vertica Benchmarks	www.vertica.com/benchmarks	See customer-submitted cost and performance comparisons between Vertica and other databases
	www.etlworldrecord.com	Vertica was used to set the data warehouse data loading world record – 5.4TB in 57 minutes
Vertica for the Cloud	www.vertica.com/cloud	Get a Vertica database instance provisioned instantly on the Amazon Cloud and use it on a month-to-month basis
Vertica Customers	www.vertica.com/customers	See who's using Vertica
Evaluate Vertica	www.vertica.com/download	Request a free evaluation copy of the Vertica Analytic Database to download and install

About Vertica Systems

Vertica Systems is the market innovator for high-performance analytic database management systems that run on industry-standard hardware. Co-founded by database pioneer Dr. Michael Stonebraker, Vertica has developed grid-based, column-oriented analytic database technology that lets companies of any size store and query very large databases orders of magnitude faster and more affordably than other solutions. The Vertica Analytic Database's unmatched speed, scalability, flexibility and ease of use helps customers like JP Morgan Chase, Verizon, Mozilla, Comcast, Level 3 Communications and Vonage capitalize on business opportunities in real time. For more information, visit the company's Web site at http://www.vertica.com.