



# Transforming the Economics of Data Warehousing with Cloud Computing

*How new frontiers in on-demand computing and DBMS technology will transform business.*

Copyright Vertica Systems Inc. November, 2008

## Table of Contents

A Bright Cloud on the Horizon...	3
5 Ways The Cloud will Transform Data Warehousing & BI.....	4
New BI technology adoption will accelerate.....	4
Organizations will conduct more short-term ad-hoc analysis.....	4
Lines of business will have the flexibility to fund more data mart projects.....	4
Data warehousing will increase within medium-size businesses.....	4
The analytic SaaS market will develop faster. ....	4
A New Analytic DBMS for a the New Frontier .....	5
Shared-nothing, massively parallel processing (MPP) architecture.....	5
Automatic high availability .....	5
Ultra-high performance.....	5
Aggressive compression .....	6
Standards-based connectivity .....	6
Case Study: Sonian Archives Digital Content in the Cloud with Vertica .....	6
Closing Thoughts & Next Steps.....	7
Getting Started with Vertica for the Cloud.....	8
About Vertica Systems .....	9

## A Bright Cloud on the Horizon...

Cloud computing is ushering in a new era of analytic data management for business intelligence (BI) by enabling organizations to analyze terabytes of data faster and more economically than ever before. The key change: cloud database software is provisioned within minutes, without data center overhead, and it's licensed on an on-demand basis.

The table below shows that organizations no longer need to justify spending hundreds of thousands of capital expense budget dollars for upfront hardware and software purchases or spend weeks waiting for hardware delivery and installation. Instead, they can sign up to tap into a computing cloud, such as Amazon's Elastic Compute Cloud (Amazon EC2).

Upon signup, Amazon EC2 completely provisions a secure, fully clustered high-performance analytic database that is hosted in the cloud and immediately ready to serve the customer's needs exclusively. They can then use it on a pay-per-use basis, usually for a monthly fee.

### Economic Comparison of In-House- and Cloud-hosted Data Marts/Warehouses

	Traditional In-House DBMS <sup>1</sup>	Vertica for the Cloud <sup>1</sup>
<b>Time to Terabyte</b>	<b>Weeks</b> to purchase, install and configure hardware and software	Immediately available - Hours from sign-up to data loading
<b>Start-up Costs/Risk</b>	Server hardware, SAN, DBMS software (>\$200,000USD)	1 month fee (\$500+)
<b>On-going costs</b>	Maintenance fees, depreciation expense, manual administration overhead	Monthly fee
<b>Data Center Overhead</b>	Floor and rack space, cooling and power costs	None
<b>Scalability</b>	Scale painfully by migrating to new hardware or DBMS. Must often over-buy up front to accommodate future growth.	Scale seamlessly, as needed, by adding instances to cluster (or removing them)
<b>Purchase Approval</b>	Complex process; funded via capital expenditure budget and requiring a 3-year ROI model	Funded via departmental operating expenditure budget requiring a 1-month ROI model

This shouldn't be confused with software as a service (SaaS) models, in which data from one customer may co-exist with data from another customer within the same application. Cloud customers are, in effect, renting dedicated servers and the people needed to house, secure, and manage them, and they have full control over server and firewall settings to ensure security.

<sup>1</sup> The Vertica Analytic Database can be deployed in-house OR in the cloud. Running Vertica in-house provides many of the same benefits described in this paper, including fast setup, higher performance and much lower hardware costs relative to "traditional" row-oriented DBMS such as Oracle, MySQL and others. Visit [www.vertica.com/benchmarks](http://www.vertica.com/benchmarks) to learn more about the Vertica performance advantage.

## 5 Ways the Cloud will Transform Data Warehousing & BI

As a new alternative to traditional, in-house data analytics infrastructure, the Cloud will transform the economics of BI and open up many new possibilities for organizations of all sizes. Cloud-based analytics should be expected to impact BI in the following five ways:

### 1. Lines of business will have the flexibility to fund more data mart projects

Because there are no long-term financial commitments required, lines of business can pay monthly cloud database usage fees out of the operating expense budgets they control rather than going through a lengthy capital expenditure approval process. Companies can fund departmental, proof of concept, and ad-hoc analytic data projects on-demand, giving them the agility to respond to BI needs faster than competitors and increase the quality of strategy setting and execution.

### 2. Organizations will conduct more short-term ad-hoc analysis

The need for data often arises suddenly, in response to new business conditions or events. The need may also last only a short time—maybe just a few weeks or months. For example, a company might need to suddenly analyze manufacturing data in the wake of a quality or safety breakdown, or it may need a new price plan in response to a new market condition. The cloud gives companies a way to respond to these requests immediately—get a data mart created in a few hours or days, have business people slice and dice to their hearts' content for as long as they need to, then cancel the cloud cluster, and it goes away with no leftover hardware or software licenses. The cloud makes it economically feasible to conduct more of these short-lived projects.

### 3. Data warehousing will increase within medium-size businesses

Despite their size, many midmarket companies have very large volumes of data they would like to analyze. Hedge fund companies with only a handful of IT people at their disposal need to analyze tens of terabytes of stock market history data to hone their trading strategies. Young bio-techs are in similar situations—they have hundreds of gigabytes of genomic data to cull through. Cloud-based analytic databases will enable them to warehouse and analyze terabytes of data even though their BI budgets and staff are a small fraction of larger enterprises.

### 4. The analytic SaaS market will develop faster

Companies that collect economic, market, scientific, and other data and then offer customers the ability to analyze it on line—analytic SaaS—will come to market faster and in greater numbers. They will be able to bring their solutions to market with much less risk and cost by basing them on the cloud during the early stages of growth. The companies can use the hundreds of thousands of dollars saved on in-house data center development to invest in customer acquisition, product development, and other market development activities. After the viability of the business model is proven, analytic data can be migrated to internal databases from the cloud if needed.

### 5. New BI technology adoption will accelerate

The cloud will become the de facto platform for evaluating new software. The cloud enables software companies to make new technology available to many more evaluators on a self-service basis. Unlike free software downloads, evaluators are spared the time and expense of finding hardware and going through installation and setup and the other tasks required to get the software up and running. As a result, the adoption of new BI software technology should increase much faster than it has in the past.

## A New Analytic DBMS for the New Frontier

In order for these pioneering analytic cloud projects to succeed -- especially as data volumes grow—they will require a database architecture that is designed to function efficiently in elastic, hosted computing environments like the cloud.

A computing cloud, such as the Amazon EC2, is composed of thousands of commodity servers running multiple virtual machine instances (VMs) of the applications hosted in the cloud. As customer demand for those applications changes, new servers are added to the cloud or idled and new VMs are instantiated or terminated.

Cloud computing infrastructure differs dramatically from the infrastructure underlying most in-house data warehouses and data marts. There are no high-end servers with dozens of CPU cores, SANs, replicated systems, or proprietary data warehousing appliances available in the cloud. Therefore, a new DBMS software architecture is required to enable large volumes of data to be analyzed quickly and reliably on the cloud's commodity hardware. Recent DBMS innovations, such as those featured in the Vertica Analytic Database for the Cloud make this a reality today, and they include:

### Shared-nothing, massively parallel processing (MPP) architecture

In order to drive down the cost of creating a utility computing environment, the best cloud service providers use huge grids of identical (or similar) computing elements. Each node in the grid is typically a compute engine with its own attached storage. For a cloud database to successfully "scale out" in such an environment, it is essential that the database have a shared-nothing architecture utilizing the resources (CPU, memory, and disk) found in server nodes added to the cluster. Most databases popularly used in BI today have shared-everything or shared-storage architectures, which will limit their ability to scale in the cloud.

### Automatic high availability

Within a cloud-based analytic database cluster, node failures and node changes can occur. Given the vast number of processing elements within a cloud, these failures can be made transparent to the end user if the database has the proper built-in failover capabilities. The best cloud databases will replicate data automatically across the nodes in the cloud cluster, be able to continue running in the event of 1 or more node failures ("k-safety"), and be capable of restoring data on recovered nodes automatically—without DBA assistance. Ideally, replication will be "active-active" in that the redundant data may be queried to increase performance.

### Ultra-high performance

One of the game-changing advantages of the cloud is the ability to get an analytic application up quickly (without waiting for hardware procurement). However, there can be some performance penalty due to Internet connectivity speeds and the virtualized cloud environment. If the analytic performance is disappointing, the advantage is lost. Fortunately, shared-nothing columnar databases like Vertica for the Cloud are designed specifically for analytic workloads, and they have demonstrated dramatic performance improvements over traditional, row-oriented databases (as verified by industry experts, such as Gartner and Forrester, and by [customer benchmarks](#)). This software performance improvement, coupled with the hardware economies of scale provided by the cloud environment, results in a new economic model and competitive advantage for cloud analytics.

### Aggressive compression

Since cloud costs are typically driven by charges for processor and disk storage utilization, aggressive data compression will result in very large cost savings. Row-oriented databases can achieve compression factors of about 30% to 50%; however, the addition of necessary indexes and materialized views often swells databases to 2 to 5 times the size of the source data. But since the data in a column tends to be more similar and repetitive than attributes within rows, column databases often achieve much higher levels of compression. They also don't require indexes. The result is normally a 4x to 20x reduction in the amount of storage needed by columnar databases and a commensurate reduction in storage costs.

### Standards-based connectivity

While there are a number of special-purpose file systems that have been developed for the cloud environment that can provide high performance, they lack the standard connectivity needed to support general-purpose business analytics. The broad base of analytic users will use existing commercial ETL and reporting software that depend on SQL, JDBC, ODBC, and other DBMS connectivity standards to load and query cloud databases. Therefore, it's imperative for cloud databases to support these connection standards to enable widespread use of analytic applications.

In summary, cloud databases, such as Vertica for the Cloud, with the architectural characteristics described above will be able to not just run in the cloud, but thrive there by:

- "Scaling out," as the cloud itself does
- Running fast without high-end or custom hardware
- Providing high availability in a fluid computing environment
- Minimizing data storage and CPU utilization (to keep cloud computing fees low)

### Case Study: Sonian Archives Digital Content in the Cloud with Vertica

Sonian is a software-as-a-service (SaaS) provider of digital content archiving. Their product, Sonian Archive SA2, archives and indexes email, instant messages and other digital content from customers' communications servers and makes it easily searchable from a Web portal. Sonian's target customers include small-to-medium-sized businesses, large corporations and government agencies – all of whom need to securely store and quickly access large data sets for compliance and reporting purposes.

In the past, customers had to invest in expensive, proprietary IT systems or expensive hosted services to meet their archiving requirements. Today, they can outsource their archiving to Sonian, for a fraction of the cost.

To provide an enterprise-class solution at an affordable price, Sonian uses the Amazon EC2 and Amazon Web Services. Amazon EC2 gives Sonian access to a cluster of virtual servers that the application can harness in real time to process data quickly for Sonian customers. The Sonian product architecture was designed to scale inside the cloud, enabling Sonian to meet customers' storage and performance demands at low cost as its business grows.

In building out its product architecture, Sonian realized that it would eventually need a database that could scale to accommodate a large number of users doing lots of queries against large data sets – without compromising on performance. The database management system would need to cost-effectively store and analyze terabytes (and eventually petabytes) of customer data. The data includes both the content (for example, the complete content of an email message) as well as metadata, descriptive information that defines the content for indexing purposes. Sonian would be storing a large amount of data for each customer; for example, for a 6,000-employee health care organization, Sonian would be archiving and managing 100 terabytes of data.

“Our whole infrastructure has been designed from the ground up to scale inside cloud computing environments, so that puts unique requirements on the database that we use,” explains Greg Arnette, Sonian’s chief technology officer. “We saw that other cloud databases wouldn’t work for what we needed – they just can’t scale, and we knew that we would encounter capacity and performance problems as our data volumes grew. Although we could have built a home-grown system – and this was our original plan – maintaining and updating this kind of system would have been too expensive and time-consuming in the long run.”

Sonian instead chose the Vertica Analytic Database for the Cloud. The Vertica Database is used to power the analytic engine behind Sonian Archive SA2. The combination of performance and cost-effectiveness enables Sonian to meet its goal of providing enterprise-class archiving at a competitive price.

“We have introduced a disruptive pricing model into a market that previously only had premium offerings,” explains Arnette. “To be competitive, however, we need to keep our costs low without sacrificing performance. The Vertica Database and the Amazon cloud compute model are the right combination to give us the scalability we need while keeping costs in line.”

To hear Greg Arnette explain in detail why and how the Sonian SA2 application was built for the Amazon EC2, visit [www.vertica.com/sonian](http://www.vertica.com/sonian).

## Closing Thoughts & Next Steps

There will surely be a myriad of DBMS offerings in the cloud as new and existing vendors attempt to capitalize on the cloud’s popularity. Many of these will NOT be marriages made in heaven. However, Vertica’s innovative DBMS software is here today and truly takes advantage of the cloud architecture in order to change the economics and the responsiveness of business analytics.

### Vertica for the Cloud Case Study



[www.soniannetworks.com](http://www.soniannetworks.com)

#### The Industry

Information Technology

#### The Application

Web 2.0 application for digital content archiving

#### The Benefits

- Fast, simple and cost-effective storage and *ad hoc* querying of terabytes (and eventually petabytes) of customer data
- Scales easily, by adding low-cost commodity servers
- Ease of integration with application’s cloud-based architectural platform

## Getting Started with Vertica for the Cloud

Vertica for the Cloud provides the “Fastest time to Terabyte.” A Vertica for the Cloud database can be ordered and provisioned on-line within minutes, and you can be loading data into it the same day. Vertica for the Cloud is licensed on a monthly basis with fees based on the amount of data you want to manage (visit [www.vertica.com/cloud](http://www.vertica.com/cloud) for pricing). The monthly fee includes:

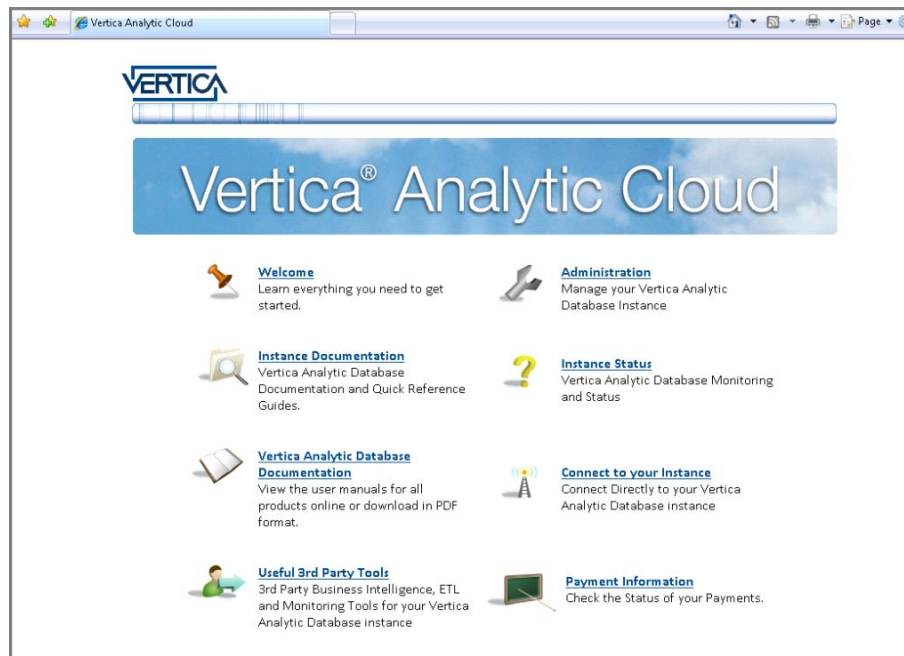
- Use of a dedicated, Amazon EC2-hosted Vertica Analytic Database instance for one month
- Amazon EC2 hardware hosting and data transfer fees
- Amazon S3 fees for data backup
- Use of Vertica management and monitoring tools via the Vertica for the Cloud console
- Vertica technical support

**Vertica Scales Limitlessly with the Cloud**

- 7.5GB RAM
- 2 x 1.7Ghz Opteron Core
- 850GB Disk
- Apache Web Server
- Fedora Core OS
- Vertica Analytic Database
- ODBC, JDBC, Python, Ruby drivers

Configuration	Raw User Data Volume
1 Node	500GB
3 Nodes	1-3TB
Additional nodes	Scales limitlessly 1TB per node

After the Vertica cluster is provisioned by Amazon EC2, simply use any SQL tool via ODBC/JDBC (or Perl, Ruby, Python, PHP) to load and query Vertica for as many months as you need. When you're done, just stop paying to use Vertica for the Cloud and Amazon EC2 reclaims and reassigns the nodes in your cluster.



If you would like to learn more about the Vertica Analytic Database for the Cloud or about building analytic database applications in the cloud, then please visit the following links:

<b>How to Build Large-scale Analytic Databases in the Cloud</b>	<a href="http://www.vertica.com/sonian">http://www.vertica.com/sonian</a>	Watch a recording of Sonian CTO Greg Arnette and Amazon Sr. Evangelist Jeff Barr explain how to architect an application for managing large volumes of data in the Amazon EC2
<b>Use Vertica for the Cloud or Learn more about it</b>	<a href="http://www.vertica.com/cloud">www.vertica.com/cloud</a>	Get a Vertica database instance provisioned instantly on the Amazon Cloud and use it on a month-to-month basis
<b>Vertica Benchmarks</b>	<a href="http://www.vertica.com/benchmarks">www.vertica.com/benchmarks</a>	See customer-submitted cost and performance comparisons between Vertica and other databases
<b>Vertica Customers</b>	<a href="http://www.vertica.com/customers">www.vertica.com/customers</a>	See who's using Vertica

## 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>.