

## Comcast Revolutionizes Network Performance Data Analysis with Vertica

Founded in 1963, Comcast is the largest cable communications company in the United States. The company serves 24.1 million cable customers, 15.2 million digital cable customers, 13.2 million high-speed Internet customers and 4.6 million voice customers. Comcast had 2007 revenues of \$24.97 billion and employs 100,000 people.

To retain and continually improve its competitive lead in the industry, Comcast puts a lot of emphasis on delivering a consistently good customer experience. In order to meet its high quality of service standards, Comcast network operators must be able to quickly collect and analyze data being generated by devices in the network.

Monitoring networks in a cost-effective manner is a challenge facing most telecommunications providers today as their network-data volumes increase. The challenge is particularly acute for Comcast, because of its size: Comcast's network has millions of components, and there are billions of metrics that could indicate a potential service interruption or other problem.

### The Application

Faced with this challenge, Senior Director of Network Operations Brian Harvell decided to build an application for doing time-series analysis of SNMP data. The application continuously samples SNMP data from network nodes and analyzes it over time to spot anomalies or trends. The concept of a time-series analysis application was familiar territory for Harvell: he had built a similar system when he was at AOL.

Comcast collects about 4 billion samples of network performance data per day (about 15 GB of data per day or 46,000 samples per second). Each sample is a 4K-8K reading that includes information about the device sampled, the type of metric recorded, a timestamp and the actual value of the performance metric. To maximize database insert speed and compression, all the data items within each sample are numerically encoded (and interpreted within reports by joining with a metadata table containing device and metric descriptions, etc.). The volume of data collected and the throughput will increase as Harvell and his team extend their monitoring coverage of the network.

### The Vertica Solution At a Glance

#### The Customer



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

#### The Industry

Telecommunications

#### The Application

- Network performance monitoring data warehouse
- Monitors millions of network devices to ensure quality of service and accuracy of capacity planning
- Inserts 46,000 new rows of SNMP data per second 24x7 (5.5TB/year)
- Sub-second query performance
- Accessed via Hyperion and Corda BI tools
- Runs on a cluster of 5 HP ProLiant DL380 servers—no costly SANs or “big iron.”
- Met rigorous fault-tolerance and recovery tests

## The Problem

Because of the volume and throughput of data, the system posed challenges when choosing a database management system.

“When I joined Comcast, I quickly realized that we needed a time-series data store to take our architecture to the next level,” recalls Harvell. “We had lots of data in different places that provided a disjointed view into network performance, but we couldn’t spend a lot of time developing a custom database internally like we did at AOL. Our problem was larger than most implementations could handle. We knew that compression would be big factor for this new system. In my previous experience, disk I/O was always a bottleneck.”

The idiosyncrasies of time-series data created extreme performance demands – a challenge even to an expert in network-performance systems like Harvell. For example, in time-series analysis, there are no off-peak periods. Whenever network operators log in to query the data, they always collect the same number of metrics, so the system must always be at peak performance. It can’t catch up if it gets behind.

Other characteristics made the data a good candidate for a column-oriented database, including fixed or known sample sizes; numeric data; table structure (the team uses schemas to divide the tables into classes of data, each with four columns); and the write-once/read-many nature of the application. Network operators are typically looking at hundreds of samples from a single node.

*“The k-safety redundancy that Vertica provides is like a dream. It replicates the data to redundant clusters for failover purposes, giving us the ultimate protection for our data. We are very pleased with the economics of replicating the data stored in Vertica relative to our other disaster-recovery initiatives.”*

Brian Harvell  
Sr. Director Network Operations  
Comcast

Column-oriented databases organize data on disk as columns of values from the same attribute, as opposed to storing it as rows of tabular records as in traditional relational databases (RDBMSs). When a query needs to access only a few of those attributes – as in Comcast’s application – it only needs to read those columns. By comparison, with traditional RDBMSs, queries have to read all values in a table, wasting I/O bandwidth and making the queries very slow.

Based on the application’s requirements, Harvell and his team decided that they needed a column-oriented database that could do the following:

- Load 50K+ samples per second
- Provide query response times of 1-2 seconds
- Provide views of annual detail, not just weekly detail
- Deliver at least 10:1 data compression
- Scale to accommodate 40+ terabytes (TB) of data using standard hardware
- Provide high availability
- Be extremely cost-effective

## The Solution

After briefly considering and rejecting an open-source tool (it was not scalable) and internal development (it would be too expensive to build and maintain), Harvell and his team chose the Vertica Analytic Database 2.0.

The Vertica Analytic Database is a high-speed, relational SQL database management system (DBMS) purpose-built for analytics and business intelligence. The Vertica Database has a shared-nothing, column-oriented architecture, and has been benchmarked by many customers as being 10x to 200x faster than other solutions. It also uses compression very aggressively, both of data on disk and on data “in motion” during queries, which further enhances query speed while enabling cost-effective storage management. For example, companies can store up to 10 terabytes of data in just 1 to 3 terabytes of disk space.

	Requirement	Vertica Performance	Advantage
<b>Query Speed</b>	1 to 2 seconds	Sub-second	10x faster than runner-up
<b>Load Rate</b>	50K rows/sec/stream	130K+ rows/sec/stream	Much faster than other column DBs
<b>Compression</b>	10:1	10.7:1	Vertica added new compression scheme during the evaluation to meet requirements for Comcast’s data set

Figure 1: The Vertica Database provides Comcast with dramatic performance advantages for time-series data management.

The Vertica Database runs on clusters of inexpensive, industry-standard Linux servers – a distributed architecture of which Harvell was a big fan. This means that when he and his team want to increase query performance, load times or capacity, they can simply add more servers – without paying a hardware or software license “tax” to Vertica.

Other features that made the Vertica Database attractive to Comcast were automated schema design and tuning, which would save on administration costs and enable fast incorporation of new data sets; and a hybrid storage architecture comprising write-oriented and read-oriented stores that work together to enable continuous loading and querying of data. Vertica’s built-in fault tolerance, which uses an approach called k-safety, provides high availability of data by maintaining duplicate copies of data across the other machines in the cluster. Should a server go down and then be restored, the system will automatically recover the restored node’s data by querying the other nodes.

Before deploying the Vertica Database, Harvell and his team conducted extensive performance testing of the cluster. The test included a snapshot of their application running on a five-node cluster of inexpensive servers with 4 CPU AMD 2.6 GHz core processors with 64-bit 1 MB cache; 8 GB RAM; and ~750 GBs of usable space in a RAID-5 configuration.

To stress-test Vertica, the team pushed the average insert rate to 65K samples per second; Vertica delivered millisecond-level performance for several different query types, including search, resolve and accessing two days’ worth of data. CPU usage was about 9%, with a fluctuation of +/- 3%, and disk utilization was 12% with spikes up to 25%. In short, Vertica worked lightning-fast without straining.

“During a try-to-break-it exercise, we were able to sustain an insert rate of more than 130K samples per second for 48 hours,” says Harvell. “We were pleasantly surprised that the Vertica Database didn’t crash during this exercise – at no time did it crash or act abnormally under this extreme load. This showed that we could drive the cluster above its capacity if we get into a situation where we need to replace some data.”

Comcast currently runs the Vertica Database on a cluster of five Hewlett-Packard ProLiant DL380 systems each with a local disk only, eliminating the need for a storage area network (SAN).

The architecture of Vertica is very similar to that of the solution that Harvell had built earlier in his career, giving him a familiar but more easily scalable foundation for his application. “We knew it would work,” he says. “The out-of-the-box loading and query performance and compression ratios that Vertica demonstrated were very good. And we were able to work with Vertica to make things even better. For example, Vertica added a new compression type for our timestamp data during the evaluation.”

Harvell continues: “The k-safety redundancy that Vertica provides is like a dream. It replicates the data to redundant clusters for failover purposes, giving us the ultimate protection for our data. We are in the process of setting up our disaster-recovery cluster and we are very pleased with the economics of replicating the data stored in Vertica relative to our other disaster-recovery initiatives.”

### The Final Word

Harvell concludes: “Vertica has been very easy to work with. The initial proof of concept was set up in just a few days. The Vertica Database runs great on the Linux gear that we have and works with our BI tools, which include tools from Corda and Hyperion. We hope to scale the system substantially, and we believe that Vertica will be able to support this growth for the foreseeable future. The Vertica support people have been very helpful and have been able to resolve any issues very quickly. Overall, we have been very happy with the Vertica platform, and I am confident that it’s a sound architecture for storing time-series data.”

### 7 Key Vertica Database Innovations

1. Column store architecture
2. Aggressive compression
3. Concurrent load and query
4. Automatic database design
5. High availability without hardware redundancy
6. Runs on commodity hardware
7. Scale by adding inexpensive servers

### Try the Vertica Analytic Database Yourself

If you would like to learn more about how the Vertica Analytic Database can help your company transform business analytics or to request an evaluation copy, please visit [www.vertica.com](http://www.vertica.com) to find out more.