HOW BUSINESS ANALYSTS BECOME ROCK STARS WITH VERTICA

Ted Cuzzillo

Datadoodle

datadoodle.com

About the author: Ted Cuzzillo is a veteran technology journalist, thought leader, and storyteller, with a series of topics going back some 30 years. For the last 11 years, he's worked in the data industry. He has written for various industry media including TDWI, Information Management, and his weblog, Datadoodle.

How Business Analysts Become Rock Stars with Vertica

Executive Summary

It’s only natural for data analysts or administrators to focus on current needs when setting requirements for new technology. Usability and features of any prospective solution are usually assessed this way.

Such a frame is far too narrow. Things change so fast today that the resulting choice will satisfy the demands at hand but is unlikely to meet new demands that arise even in the near future. Suddenly, the once shiny solution has to be replaced, upgraded, handed a crutch. While any number of narrowly focused tools on the market might handle current jobs, best equipped technology to meet today’s and tomorrow’s needs is mature, tested, and solid.

The wise way to evaluate technology allows for unforeseen needs. That wider frame leaves room for other factors to reveal their true importance. Depth of analytics functionality, scalability, the future expansion of operational areas, the number of users supported, and the type of analytics they run usually do evolve, enlarge, and even inspire unforeseen applications that deliver new business value.

This is where the Vertica Analytics Platform excels. This is why data analysts use it and even love it. Some even speak of “shock,” as one analyst put it recently, at Vertica’s speed. They find efficiencies in its administrative ease, and find new analytical modes they didn’t even know existed.

The Known and the Unknown about the Vertica Analytics Platform

The story goes that a Vertica sales rep was presenting an update at a customer site, and he mentioned a built-in analytical function—perhaps time series, or geospatial, or machine learning. No one quite remembers what. The rep does remember, though, a reaction from the back of the room: Two data analysts, listening intently, turned to each other and slapped high fives.

This is actually a common story with Vertica. The analytical database's better known speed and versatility tend to upstage its analytical functions, performance, and other capabilities. When analysts discover new analytical capabilities—whether the users are first-time users running traditional analytics during a proof of concept or long-time users realizing Vertica includes time series analytics—these capabilities bring surprise and delight.

Vertica's surprises include these features:

- Built-in geospatial and time series analytics.
- Machine learning
- Projections, similar to materialized views, are copied, compressed, and highly optimized data that's stored to speed up particularly long-running queries.
- Workload management that recognizes various priorities among analytics jobs. This lets some queries run ahead of others. Some queries use fewer resources and take longer while others need to run fast.
- Cascading pools account for poorly formulated queries, where queries can be bumped down to a lower resource pool if they take too long.
- To help satisfy GDPR and other new regulations governing personally identifiable information (PII), column-level encryption supports multiple teams in which access varies among users.
- A management console that lets administrators visualize and manage queries, pools, and nodes.
- A security module that runs Apache Sentry, Kerberos, and Voltage SecureData .

The discoveries are best told in stories from the data analysts themselves.

> *The wise way to evaluate technology allows for unforeseen needs. That wider frame leaves room for other factors to reveal their true importance. Depth of analytics functionality, scalability, the future expansion of operational areas, the number of users supported, and the type of analytics they run usually do evolve, enlarge, and even inspire unforeseen applications that deliver new business value.*

Analysts’ Stories

Ameripride, Industrial Laundry

Industry: Services

Business Analyst Value:

- Capability to generate in-meeting reports during the meeting
- Visibility into previously unseen revenue opportunities
- Handle RFID data too big for legacy solutions

Some people look at industrial laundry and see just uniforms, linens, and floor-mats. Anthony Ordner, director of information management at Minnetonka, MN-based AmeriPride Services, sees data.

Much of his data combines traditional data warehouse data from sales and logistics with specialized data from embedded RFID tags on the truckloads of textiles. Every single day, bundles go out, come back, get washed, and go out again—generating data on washes, degree of wear, and countless other metrics. Along with these metrics, customer data helps the business run.

Some people look at industrial laundry and see just uniforms, linens, and floor-mats. Anthony Ordner, director of information management at Minnetonka, MN-based AmeriPride Services, sees data.

In an industry in which the key competitive differentiator is automation, enabled primarily by data analysis, understanding that data is crucial.

Now that Vertica is the “backbone” of their data capability, Ameripride’s ability to see and analyze the data, and the operation as a whole and in detail, has taken a huge leap. “It’s the ability to slice and dice by customer vertical,” said Ordner. “We want to understand the typical customer metrics. Why did people stop using our services? Why did they start? How are the product categories trending?”

CRUCIAL MEETING

In an interview at Vertica’s Big Data conference, Ordner recalls the moment Vertica proved itself to Ameripride’s leadership. It was in a meeting, soon after his team had deployed Vertica. “I had three years’ worth of data showing on my laptop with Tableau,” he recalls. “We were going over company sales and renewal performance, and everybody was looking at this high-level data. I was sitting off to the side.”

He offered an insight based on the data he saw—insight that related directly to the point under discussion at that moment. With the company’s previous systems, he probably could not have executed the report at all. “There were a lot of ‘ahas,’” he recalls. “And then they asked how I had the data. ‘What are you doing? What did you just do?’ Before this, I had never been able to do any type of analysis on the large data sets to see trending and results in such an interactive way. It’s always been aggregates.”

Vertica’s ability to handle huge data sets, and to do it fast, allowed analytics for the first time to become part of the conversation. It was a breakthrough for Ameripride.

VERTICA TAKES HOLD

For Tony, this was quite a career booster. Directors, VPs, higher level executives were so excited that they told their teams and got Ordner’s team to come train them. The first team to adopt the new platform was direct sales. The analytics power user on that team had had to make do with made-up numbers or approximations.

But, with Vertica, she had the real numbers—with a direct result: smarter contract renewal.

The analytics showed that total pounds of laundry keep rising, and, as a result, rising revenue should have resulted. Instead, revenue was flat. Now, with the real numbers in sight, they found a main cause: ill-guided pricing on contract renewals. At renewal time, many customers asked for discounts, and getting them. “Each was treated on a one-off basis. No one could see the trend. No one saw the whole picture. We were giving price concessions, sometimes 25 or 30 percent. That’s a problem.”

They could now see data at a high level, with Tableau on the front end, and apply it on an action level with the same data stack. With that, average renewals improved from around 20 percent to trending around 7 or 7.5 percent.

Nothing is more empowering for an analytics team than providing metrics that impact revenue directly.

PRICING MODEL MADE EXPANSION AFFORDABLE

Vertica's pricing, based on quantity of data, made expansion relatively inexpensive. "We can accelerate up or out," said Ordner, "just by throwing hardware at it."

Their former platform, Oracle, was priced by the core. For a mid-size company, which they were before the merger, this was prohibitively expensive. Ordner recalled, "We were very limited. We couldn't afford to keep throwing CPUs at it."

For analysts and the technology team, this change also represented cost savings to the organization. Lower licensing costs, fewer computing nodes required, and better access to data without the need for aggregation were all strong factors in lowering costs on IT.

COST REDUCTION: GARMENT RECOVERY

Sometimes, garments don't come back. The data was always available, since RFID tags track outgoing and incoming garments. But the sheer volume made garments hard to track. With Vertica, Ameripride could track it all. "That's another big chunk of change we're looking at," Ordner said.

LETTING ANALYSTS BE ANALYSTS

Analysts love Vertica for one simple reason: It allows them to do more of what they're trained to do, what they like to do, and what's most valuable for the company. Ordner said, "It lets them find those great patterns they just didn't have time to do."

TravelBird, Inspired Travel

Industry: Online Travel

Business Analyst Value:

- Performance sufficient to power recommendation engine.
- Easier administration than previous solutions. No DBA required.
- Fast queries can run at high priority, less important queries can be given lower priority.

Many Americans on vacation in Europe have gone hoping for more than trains, planes, and meals. They want inspiration, and that's why some consult with TravelBird. The trick is that inspiration is highly personal. To ensure that recommendations inspire and not bore individuals within the target market—which varies from around 50,000 people to up around 500,000 customers—TravelBird has created a sophisticated recommendation engine.

Recommendations start with an individual's browsing history and past travel. Vertica chews on more than a dozen sources of data. Customer preference, product similarity, "hotness" based on trending or outperforming destinations, diversity for each recipient based on recommendation confidence, lifecycle state, and yield optimization all have a part.

From the machine's cut, human agents make a further selection. Even the customer's reactions to the final portfolio gets poured back in to train the machine.

VERTICA FOR EASE OF ADMINISTRATION, TUNABILITY, CONCURRENCY, FINE-GRAIN MANAGEMENT

Rob Winters, the TravelBird head of data science, uses Vertica for two main reasons: ease of administration and performance tunability. Also, TravelBird relies on Vertica's machine learning for continuous tuning of selection models.

TravelBird uses Vertica for a wide range of jobs, from very large and long-running to short, ad-hoc work.

Among all those, Vertica manages concurrency. Other platforms—Winters mentions Amazon RedShift in particular—have made it an issue, unable to tell some queries to fall into lower-priority resource pools and balance memory and CPU allocation.

"I'm actually able to do fine-grain management of resource pooling and priorities," he said. He can have 70 queries running concurrently, some with huge requirements and some with short-burst analytics. "That's a huge win for me and my team because we don't have to sit and wait for five minutes for the query to start."

New York Genome Center, Seeking Cures for Complex Diseases

Industry: Medical Research

Business Analyst Value:

- Sophisticated, complex analysis done quickly
- Easy discovery of cohorts with data at high volume, sometimes rows in the tens of trillions
- Statistical validity with a wide variety of data analyzed

Alzheimer’s, asthma, diabetes, Parkinson’s, and “Lou Gehrig’s disease” are all thought by medical science to arise from a combination of genes that have gone bad. Finding those failed genes involves heavy data analysis with a database to suit.

That analysis, on the Vertica Analytics Platform, is what researchers at the New York Genome Center have undertaken. The Vertica Analytics Platform has been helping them analyze terabytes of genome data quickly and efficiently.

The raw data amounts to about 12 terabytes daily, and it could double by the end of 2018. She told Dana Gardner on Briefings Direct podcast, “In a lot of environments, you start with the raw data, you analyze it, and you cook it down to your answers. In our environment, it just gets bigger and bigger.”

One project ingests data streaming from devices. Soon, cardiac monitors, glucose monitors, and other wearable devices will likely add to that flow. Already, patient data on rheumatoid arthritis comes from smartphones.

It’s a wide variety of data to ingest. Full medical records, genomic data loads up along with the streaming data. Some current studies will need 100,000 patients for statistical validity, which requires all the data also to reside in one place.

METHODS

Machine learning and bayesian statistical analysis extract information from the large chunks of incoming data. The sizes can be from 150 to 500 gigabytes to a terabyte or more each. Correlations of disease with DNA variance and mutations provide a first look. The data then flows into a database with all the other data that’s already accumulated for researchers. Bloom said, “We want to let them find more data like the data they have, so that they can get statistical validation of their hypotheses.”

The more data researchers have, the more easily they find patient cohorts. How likely is it, they ask as they evaluate each cohort, that a given genomic variant will correlate with a given disease? Or, if the disease is already present, how likely is it that the variant is present?

“You can only perform such analysis,” Bloom said, “if it’s easy to find all of that data together in one place in an organized way.”

Only a flexible database can connect data from medical records, for symptoms and diseases, with DNA data, RNA data, epigenetic data, with microbiome data.

THE RIGHT DATABASE

The right database wasn’t easy to find. “We were looking for one that could handle tens of trillions of rows without falling over,” she said. The database has to let them easily change and add new kinds of data. She said, “We’re always finding new kinds of data to correlate.”

Other priorities also loomed. Because the center deals with healthcare, special attention went to privacy, governance, and auditing. Because the center is a non-profit, it was sensitive to cost.

The New York Genome Center’s questions are hard to answer—but they’re even harder to bear if there are no answers. “Think about the impact of these answers on all of us,” she said. “If we can find the molecular causes of Alzheimer’s, for example, it could lead to treatments or prevention and all of those other diseases as well.”

hMetrix, Data Solutions for Healthcare Organizations

Industry: Healthcare

Business Analyst Value:

- Performance sufficient to power near real-time recommendation engines
- Vastly simplified administration over previous solutions minimizes the time needed from DBAs and ETL specialists

- Streamlined queue prioritization means fast queries can run at high priority while less important queries are pushed down appropriately

For Zacharia Mathew, vice president at hMetrix, Vertica's advantage is plain to see. For hMetrix, which provides a wide range of data solutions to healthcare organizations, Vertica's storage optimization means that Mathew and his team don't have to start every job worrying about how to store and optimize ever-increasing volumes of client data. They can directly load the data and proceed straight into data analysis and insights, which is what clients expect from hMetrix.

Data volume can run into billions of claims, requiring terabytes of space. hMetrix tried several databases before Vertica but found that each required considerable time planning and optimizing data for speed. Without Vertica, they began each new project with basic questions: Can they store the data? Is storing it all in a denormalized table feasible and optimal for analysis? How can they store the data so that an analyst can examine it quickly and efficiently?

The considerable time spent answering them is usually an uncompensated expense. "If I'm the client," said Mathew, "I care about what's happening to the patient, how my hospital is running, how are my physicians performing against the benchmarks?" he said. "The earlier we get to those questions, the sooner the client is happy, and the sooner I'm happy."

The company started using Vertica in 2009. Performance was so fast, he said, "We could not believe the improvement." Query response was in some cases 70 times what they'd seen with other solutions. At meetings with new partners, clients expressed awe.

hMetrix was also pleased that Vertica required so little administration. "hMetrix' core competencies are data science and the translation of data into insights—the less we need to worry about infrastructure and database maintenance, the better," he said. "In most cases, improving responses from five seconds to four seconds isn't worth the trouble. We're getting an optimized, out-of-the-box response in a few seconds, which makes fine tuning unnecessary."

The ease with which hMetrix can maintain databases containing hundreds of terabytes of data is a tremendous advantage. Competitors typically deploy numerous individuals before an analyst even sees the data: ETL specialists make modifications to load the data; DBAs check and tune the database; and analysts may have to return to the DBA throughout the analysis. At hMetrix, the analysts and data scientists themselves can load the data in a secure and segmented routine and begin their analysis immediately.

One more thing he likes: Vertica is extremely reliable. "Its stability and dependability have been a constant reassurance to us and our clients over the 9+ years we've been using it."

Nimble Storage, Big Storage Managed with Big-Data Analysis

Industry: Data Management

Business Analyst Value:

- Seconds, not hours, for answers from petabytes of storage
- Easy administration of data storage based on large volumes of operational data
- Data analysis with mature, known technology

Nimble Storage analyzes big data to report on and manage its big-data storage products. The company offers SSD storage and traditional hard disk storage, and many combinations of the two. With data analysis, it gets the best of two worlds—the economy of inexpensive but slow traditional disks and the performance of fast but expensive flash drives. Nimble's analysis shows it what data needs to be accessed and what data probably won't be called on often. With that, Nimble knows how much cache a customer needs.

"We recognized that if we could collect enough operational data about product performance," explained Larry Lancaster, the former Nimble Storage chief data scientist (now founder of Zebrium), "and then get it back home for analysis, we could dramatically reduce support costs."

Without such fast analysis, managing customers' storage would be far more difficult. Instead of almost instant analysis, they might have had to simulate performance in the lab. "That's a very labor-intensive, slow process with very little data to base the decision on," he said. With Vertica, Nimble sees in a near real-time workload distribution in the field. They also see how customers access storage.

"We have a very tight feedback loop," he said. "I think we have a better understanding of the way storage works in the real world than any other storage vendor, simply because we have the data," said Lancaster.

PERFORMANCE BENEFITS

“Internally, we’re using Vertica, just because of the performance benefits,” he said. One particularly large query looked at certain aspects of latency over a month, across the entire installed base, to understand a little bit about the distribution, depending on different factors.

They first ran the query in Postgres. Results came back 12 to 24 hours later, depending on the server’s load. Then they ran the query on Vertica, and it took from three to seven seconds. Smaller queries get sub-second latencies. On big ones, sub-10-second latencies. “It’s absolutely amazing. It’s game changing.” With that kind of response time, analysts could stay at their desks, he said, and iterate. There was no longer any need to start a batch and let it run overnight. “Data scientists tend to be impatient,” he said. “They’re highly paid people. You don’t want them sitting at the desk waiting.” “It’s hard to find a database that similarly capable,” he said. “That’s how I think of Vertica.”

COMING BACK TO THE DATABASE

Would Hadoop do better? “You have a bunch of guys with their red hats on for Hadoop,” he said, “and the next thing you know they’re building their own SQL database on it. I don’t know why you’d do that. “The bottom line,” he said, “is that the market no longer needs just a cheaper data stack and a pretty GUI. It needs platforms and components suitable to build critical applications on top of, and to integrate across business systems, cloud or otherwise.”

Hadoop enthusiasts will be back to the database, he predicts. “You know, there are databases out there like Vertica that are MPP scale-out,” he said. “And the only other thing that’s missing is having structured data to work with. To me, that’s why Vertica’s special. It does everything it can to enhance performance to get answers out of petabytes of data.”

Conclusion

One thing often goes unsaid in all the acknowledgement of Vertica’s speed, data management, and built-in analytical functions. It’s that Vertica lets analysts be analysts—doing the work they trained for, the work they like, and the work with the most value.

With Vertica, analysts go straight to data analysis without worrying about where to store the data, or without jumping to another tool for time series, geospatial, or other analysis, or timing a complex query to run overnight instead of right now. Vertica just handles it for them.

The analyst stories say it best:

- Out of the box, Ameripride tightened up its contract renewals when data revealed the one-off approach to renewal discounts. Average renewal discounts went from around 20 percent to about 7 percent.
- Ameripride’s recovery of expense for unreturned garments improved with the ability to use all the RFID data received.
- TravelBird improved its recommendation engine.
- Complex diseases may finally find a cure at the New York Genome Center, running Vertica.
- hMetrix has eliminated the need for a full-time DBA because of Vertica’s easy administration.
- The data scientists at Nimble Storage got to do what they wanted to do: derive insight from data.
- At hMetrix and Ameripride, analysts could at last stop waiting for data to process. With Vertica’s speed, they could do more of what they’d been trained to do, and where they delivered the most value. They analyzed data.

As Tony Ordner showed in that meeting, Vertica makes analysts into rock stars.

Learn why thousands of other leading data-driven organizations rely on Vertica at:

www.vertica.com