ActualTech Media

MICRO FOCUS®

IoT: From Edge to Core to Cloud

Joey D’Antoni

CONTENTS

IN THIS PAPER

With the number of edge computing devices continuing to grow exponentially year over year, designing IT architectures to capture, process, analyze, and derive deeper insights from that data will be one of the defining problems of the next decade.

This tech brief discusses how HPE Edgeline Systems, HPE MapR, and Micro Focus Vertica enable organizations to process massive amounts of data at the edge, as well as explore and gain deep insights from petabytes of data.

INTRODUCTION

Whether gathering user data from mobile devices or sensor data from manufacturing lines, much of modern computing has shifted to the "edge." The edge is a computing term, but it's also the location where transactions take place—a retail sales floor, your home, a power plant, an oil rig, or an airplane, but not in a centralized data center as in the past.

> The technology that brings intelligence to the edge encompasses three "C's"—connect, compute, and control.

The technology that brings intelligence to the edge encompasses three "C's"—connect, compute, and control: The devices involved are nearly always **connected** to the Internet, allowing them to transmit data to a central analytics system; they all have microprocessors, so they can provide **computing** power to enable access to applications, make decisions, and filter data; and that computing power lets these devices **control**—change settings to manage specific scenarios or to orchestrate certain actions. Such devices are connected to sensors and actuators to control things—from smart lightbulbs in your home to a turbine in an electric plant.

Edge computing led to a design pattern called Lambda architecture, which manages large volumes of data by leveraging both stream and batch processing methods. The streaming layer is used to process data in real time and is useful for identifying anomalous values, like a temperature being out of range, and then either taking an action or firing an alert. The batch layer stores the large volumes of data coming from devices and provides the computing power needed to drive analytics. Whether for data science or more traditional data warehouse queries, the batch layer needs to be able to rapidly load and query immense amounts of data.

In addition, the edge allows organizations to capture insights from new sources located closest to customers and operations. The need to quickly turn data into insights and actions necessitates a shift in how data is captured and processed locally. This in turn creates the need within applications to drive increased intelligence and automation with artificial intelligence/machine learning capabilities.

Luckily, there's a platform that can meet all of these challenges. It's called Vertica and, as a scale-out columnar database, it can both ingest large volumes of data and serve massive queries very quickly while also providing faster predictive analytics through in-database machine learning.

THE VERTICA PLATFORM

Legacy databases weren't designed to deal with the volumes of data modern analytical systems have to handle, and they require extensive tuning, optimization, and proprietary hardware in order to meet the pace of business. Vertica takes a different approach. It's what's known as a columnstore database, which offers a number of benefits for large-scale analytic workloads. The data is stored in columns, rather than in the traditional mixed data pages of a legacy relational database. This allows for much higher degrees of compression, which greatly reduces the storage space required for the data, and that reduced space in turn allows the data to be processed much more quickly. The massively parallel engine enables scaling out the workload across multiple servers, and it means data can be stored in an ordered fashion that improves performance on frequently run queries. Some public cloud platforms offer similar architectures running in a Platform-as-a-Service (PaaS) model, but the costs of the service can grow exponentially as consumption grows.

> Vertica can ingest data rapidly, with enough performance to let analysts explore data interactively and enable everyone to make data-driven decisions.

Vertica also supports connections to a wide variety of "Big Data" platforms, such as MapR (now part of HPE), Cloudera, and Apache Spark. This allows Vertica to be an analytics hub, modernizing data architectures. Vertica ships with R and Python, as well as with prebuilt machine learning options, within the database, allowing your data science to take place next to your data, which reduces the complexity of data movement and improves overall system performance. Vertica also ships with time-series and geospatial data tools, allowing you to quickly build reports on advanced data structures.

This architecture means Vertica can ingest data rapidly, with enough performance to let analysts explore data interactively and enable everyone to make data-driven decisions. Vertica customers report that compared to legacy database platforms, they see as much as a 50x boost in performance, so a report that used to take an hour now runs in less than two minutes. Vertica also applies deeper analytics capabilities that support the scale and performance of today's speed of analytics, thus creating insights and driving actions at the edge.

And, best of all, Vertica is an enterprise-ready ANSI-SQL analytics solution that's easily deployed by existing IT staff and doesn't require a highly skilled team of data scientists.

IMPLEMENTING THE INTELLIGENT EDGE

Edge implementations typically consist of a variety of hardware devices, leading to difficulties in monitoring, connectivity, and standardization. HPE Edgeline systems meet that challenge with three points of convergence of IT/OT systems that enable building an intelligent edge: deep analytics/compute, data acquisition/control, and enterprise management. The deep compute engine handles data aggregation and preparation and provides artificial intelligence by running machine learning models at the edge, in a hardware package designed to work in hardened environments with extreme temperatures and conditions. This package can be combined with systems likes Spark or Kafka to do stream processing, and you can use Vertica on the Edgeline to handle edge use cases like meter readings.

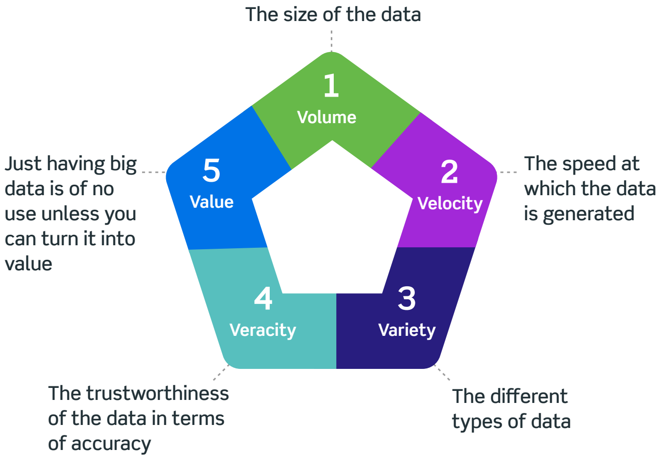
While these tools are very powerful, various customers have analysts who have many years of experience working with business intelligence platforms and data warehouses, however, lack experience working with open source tools like Spark or Kafka, which require a different set of development skills. While most data scientists possess these skills, many organizations are still growing that skillset. Using Vertica as an edge tool allows for the BI analysts to utilize their existing skill set on edge data.

> Bringing intelligence to the edge can reduce both network and decision-making latency, shrinking the amount of network bandwidth required.

Bringing intelligence to the edge can reduce both network and decision-making latency, shrinking the amount of network bandwidth required. It can also protect data against network threats and data corruption from intercepted traffic or broken connections.

VERTICA AND HADOOP, BETTER TOGETHER

Big data workloads are defined by the five "V's": velocity, volume, value, variety, and veracity. Velocity and volume are implied with edge workloads—there's a lot of data and that data is typically "as it happens" data, so it can flood the target system in no time. The value proposition is specific to the data as it relates to your business, and variety can refer either to the type of data (video, audio, documents, meter readings) or the business function of the data (supply chain versus sales, for example).


The size of the data
1
Volume
2
Velocity
The speed at which the data is generated
3
Variety
The different types of data
4
Veracity
The trustworthiness of the data in terms of accuracy
5
Value
Just having big data is of no use unless you can turn it into value

Source: Science and Education publishing (SciEP)

Hadoop is the most broadly adopted big data system, but there are challenges: Many organizations that made investments in Hadoop failed to see any business benefit, because performing analytics on massive volumes of unstructured data isn't easy. The tools used to analyze data in Hadoop were difficult for developers and analysts, and the methods of data extraction differed from normal SQL queries.

> The only way to protect data end-to-end is to put the encryption controls within the data itself using format-preserving encryption (FPE), which persists the protection as the data moves from database to application.

But there's enormous business value in capturing large volumes of data and running fast analytics for deeper insights at the edge. Most modern implementations use a data lake methodology to capture raw data in its original format. Vertica can help accelerate your time to value with this data by providing an interface that enables access to scalable ANSI SQL. Combining Vertica with Hadoop ecosystems like HPE MapR Data Platform builds a complete, integrated analytics solution that offers fast native SQL on Hadoop without the hassle of connectors. And with a single data set from Hadoop and Vertica, analysts avoid writing map reduce jobs and using unfamiliar tools, and simply write queries or use a reporting tool to analyze data.

Vertica also supports business intelligence solutions such as Tableau, QlikView, and MicroStrategy.

SECURING YOUR DATA WITH VOLTAGE SECUREDATA

Given the volume of data that can be stored in a system like Vertica, it can become a target for intruders. The common solution in databases has been to simply encrypt the data in the database, but that means the data has to be decrypted for use in other systems like business intelligence tools. The only way to protect data end-to-end is to put the encryption controls within the data itself using format-preserving encryption (FPE), which persists the protection as the data moves from database to application. Voltage SecureData offers datacentric security that includes FPE and works with big data platforms such as Vertica and Hadoop.

Let's look at a quick example that shows how FPE—and Voltage SecureData—can help. Suppose you want to encrypt U.S. Social Security numbers (SSNs), which are stored in the format nnn-nn-nnnn. If you send SecureData a value, such as 123-45-6789, to encrypt using the SSN format, SecureData then replies with, for example, 452-32-2323, which means the encrypted values don't require any changes to the underlying data structure. Not only is this much easier to implement, it means you can expose "real" values for a portion of the data, such as showing the last four digits of the SSN to a customer service representative.

THE BEST DESIGN FOR YOUR ARCHITECTURE

The number of edge computing devices continues to grow exponentially year over year, and designing IT architectures to capture, process, analyze, and derive deeper insights from that data will be one of the defining problems of the next decade. Using HPE Edgeline Systems for your edge analytics, in conjunction with Vertica, lets you process massive amounts of data at the edge. The advanced analytics offered by Vertica enables deep exploration of the data, and when combined with HPE MapR Data Platform, allows you to derive insights from petabytes of data. Finally, you can keep all that data secure using Voltage SecureData, which can directly integrate into your Vertica environment and encrypt your data with virtually no changes to your data structure or reports.