ActualTech Media

VERTICA

Future State of AI Architecture: A Powerful Combination of Platforms

Joey D'Antoni

CONTENTS

IN THIS PAPER

While your analysts and developers may have a great deal of experience with a traditional data warehouse platform, moving to new paradigms like predictive analytics and new data platforms can be challenging.

This paper will describe a powerful AI platform combination that leverages HPE compute and storage, the HPE MapR™ Data Platform, and HPE Container Platform, as well as Vertica's database to provide the quickest possible time to insights.

Data science, machine learning (ML), and artificial intelligence (AI) are all the buzz in business technology. Various vendors have made efforts to reduce barriers to entry for using these technologies or integrated them into existing projects. While this can be beneficial for helping users organize their calendars, or even go as far as helping your customers via a chatbot on your website, the types of ML/AI projects that really drive business value are much harder to achieve.

There are several reasons for this, but the first is that the toolset in this space is massive—there are open source tools, SDKs from vendors like Microsoft and Google, independent models and papers, and curated ISV solutions. Data wrangling from a variety of internal and external data sources often requires a combination of system administration, development, data engineering, and mathematical skills that are a challenge to find in a team of people, much less one or two hires. Additionally, to hire these rare individuals, you're likely competing against large software firms, financial services companies, and private consulting firms—all who pay relatively high salaries.

Another option is to try to elevate the skills of your business intelligence team from retrospective analysis to a predictive model. This approach is particularly effective when the questions you're asking of your data are straightforward, like changing production volumes based on historical trends. But those teams may lack the skills needed to integrate more advanced external datasets and build ML pipelines to improve the accuracy of models.

Vertica for Data Science

Gartner Inc. reported in January 2019 that 80% of analytics insights projects will not deliver business outcomes through 2022, and that 80% of AI projects will "remain alchemy, run by wizards" through 2020.[1] Key to building a successful data analytics project is getting the right data sets to the project, and being able to quickly perform analysis on the data.

[1] Source: Gartner Blog Network, "Predicts 2019: Analytics and BI Solutions," https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/ (Jan. 3, 2019)

Even a traditional data warehouse project involves pulling together data from a wide variety of transactional systems. In most warehousing projects 70% to 80% of project effort is spent on building extract, transform, and load (ETL) processes, which convert the data into a suitable format for analysis. In a modern analytics project that combines traditional business intelligence (sales, inventory, production, and so on) with mobile app, clickstream, and social media, data is far more complex, especially when trying to integrate with modern data storage technologies such as object stores or Hadoop. While open source systems are robust for storage and data analysis, integrating them into an enterprise data landscape is challenging because of varying connectivity and programming language options.

> Data wrangling from a variety of internal and external data sources often requires a combination of system administration, development, data engineering, and mathematical skills that are a challenge to find in a team of people, much less one or two hires.

Vertica for SQL on Apache Hadoop provides full ANSI-SQL functionality for the data lake. This means your analysts can quickly query data in the data lake and the data warehouse. You can load structured data directly into Vertica or use external tables to query data across Parquet, ORC, JSON, and other data formats on HDFS.

BUILT-IN MACHINE LEARNING

One of the benefits of using Vertica for ML projects is the in-database framework for advanced analytics and ML. This allows your analysts to quickly analyze your data using familiar SQL algorithms and combine them with ML toolsets like R and Python. Vertica's built-in ML functions include linear and logical regression, k-means clustering, and Bayesian analysis, among others. This enables you to prepare your data for normalization, outlier detection, and sampling, and to create, train, and score those models using SQL skills commonly found in enterprise organizations.

The power of Vertica's column-oriented Massively Parallel Processing (MPP) architecture means you can execute queries over hundreds of terabytes of data very quickly, on a single platform. Developers also have the option of creating functions in R, Python, Java, and C++ for custom application. Unlike many open source tools, Vertica has full commercial support and provides regular upgrades and updates to its functionality. This delivers low management overhead for your data analytics platform.

HPE Container Platform, HPE MapR, and Vertica

Hadoop is an established platform. However, it lacks many enterprise management resource controls and generally requires a team of skilled administrators to secure and manage your clusters. Historically, multi-tenancy has been a challenge in the environment, as well. The HPE MapR Data Platform offering has always been at the forefront of reducing these challenges to enterprises by minimizing the management effort through enhanced tooling and easier enterprise security integrations.

The HPE Container Platform is an enterprise-grade container platform that supports both cloud-native and non-cloud-native monolithic applications with persistent data. It includes innovations from HPE's recent acquisitions of BlueData and MapR, together with open source Kubernetes for orchestration. BlueData has a proven track record of deploying non-cloud-native AI and analytics applications in containers and MapR brings a state-of-the-art file system and data fabric for persistent container storage. Now enterprises can extend the agility and efficiency benefits of containers to more of their enterprise applications—running on either bare-metal or virtualized infrastructure, either on-premises, in multiple public clouds, or at the edge.

MODERN DATA FLOW

Vertica is typically used as a big data analytics engine because of its high performance. However, with the massive data volumes coming from Internet of Things (IoT) devices and other streaming data sources, many organizations land the raw data into a data lake, which represents a lower-cost storage platform for the undistilled data. Kafka, a streaming query engine, is also commonly used in this architecture to

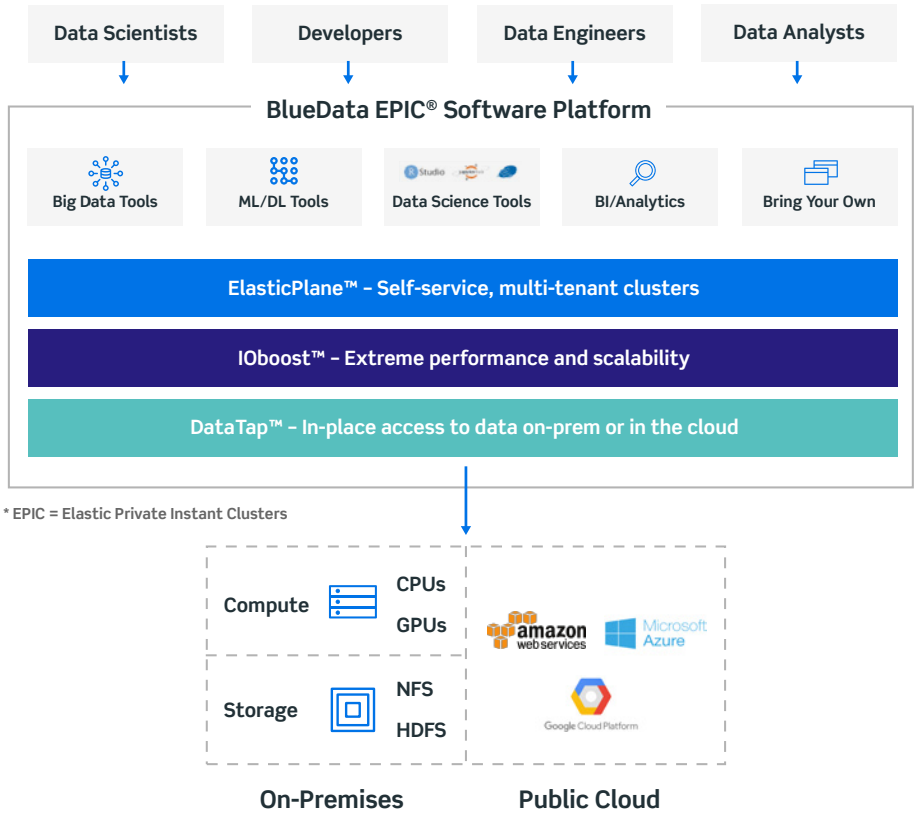

Figure 1: BlueData EPIC—a component of HPE Container Platform.

provide alerts on out-of-bound values in real time. A good example of this might be data coming from an airplane engine—most of the data is uninteresting, except for longer-term analysis. However, high temperature might require action as soon as the plane touches the ground.

The full data set will typically live in HDFS on Hadoop. The HPE MapR Data Platform provides some major enhancements around security and performance. HPE MapR's approach to security is that the product is secure out of the box, offers ends-to-end encryption, and provides a unique data governance and lineage solution, allowing you to track the data in your environment. HPE MapR also offers enhanced performance over open source Hadoop solutions by using the MapR XD file system that accesses storage directly.

BLUEDATA EPIC—A COMPONENT OF HPE CONTAINER PLATFORM

The container-based BlueData software platform is the foundation of the HPE Container Platform. With BlueData EPIC software, you can create distributed environments for ML, data science, and analytics in minutes rather than months. You can offer a self-service experience with the data and tools that your data science teams need, while providing enterprise-grade security and reducing costs. The additional value BlueData EPIC brings to the table includes IOBoost™, which is an application-aware caching service to reduce the cost of IOs against a large data set, DataTap™, which allows in-place access to data reducing costly data movement inside of a cluster, and ElasticPlane™, a management control plane that allows for multi-tenancy, and even multi-cloud deployment and management. ElasticPlane allows for automated integration with your Active Directory or LDAP solution, and management of your containerized AI and big data solutions can be easily secured and managed (see **Figure 1**).

HPE GREENLAKE

Choosing the right amount of computing resources and software for your growing data environment can be a challenge. Frequently, firms will underestimate the amount of resources they need as new data sources come into projects and require more software, storage, and compute to meet the demands of the business. In many cases, the business moves data sources into new computing paradigms over time and the IT organization can remain over-provisioned for years. The HPE GreenLake model treats software and infrastructure, both compute and storage, like public cloud resources. Software licenses are scaled with an as-needed model. GreenLake right-sizes the solution from Day 1 and adds a buffer layer of hardware that isn't paid for until it's consumed and grows as the consumptions increases. This allows for nearly on-demand scaling of software, compute and storage resources.

Derive Insight from Large Volumes of Data

While big data analytics is a rapidly evolving space, the technology stack has settled down in recent years. Most organizations are using some combination of Hadoop and an in-memory database engine like Vertica to derive insight from large volumes of data. This gives a couple of benefits—the first is that developers and business analysts who have vast experience with the SQL programming language can quickly get up to speed. The second is that the reduced cost for storage in a platform like Hadoop can be realized for a significant portion of the data store.

This architecture makes data and ML options available outside the elusive realm of a few data scientists or statisticians that you have on staff. It also allows for you to use your favorite business intelligence visualization tools like Tableau or Qlik to get your data answers in front of business users.

The Vertica/HPE MapR/HPE Container Platform combination allows businesses to have a production, Internet-scale big data solution very quickly by reducing the integration challenges that many open source projects face. When combined with the ease of deployment of HPE GreenLake, you can derive new business insights in minutes instead of months.

Things to Think About

Where is your enterprise in terms of analytics maturity? Are you just getting started? Or do you have a robust data warehouse that you'd like to augment with outside data sources, or even social or clickstream data?

Visit the HPE big data analytics solutions website and the Vertica website to discover how you can transform your business from edge to cloud and put your data to work.