ActualTech Media

VERTICA

Modernizing the Enterprise Data Warehouse: How to Level-Up Hadoop Environments with Vertica

Joey D'Antoni

CONTENTS

Introduction 2

Greenfield Analytics 2

Siloed Data—All the Hadoop Clusters 2

Overloaded with Data Capacity—My Cloud Bill Is Too High 3

Changing Analytics Requirements 4

Vertica and Hadoop—Better Together 4

Procurement Options 4

IN THIS PAPER

Many organizations face challenges as they embrace digital transformation and move toward data-driven decision making. In order to produce quality, real-time insights they need to have an analytics platform that can be broadly adapted and used across their organization. In this paper you'll learn about different approaches businesses can take to become more data-driven, and how Vertica and HPE MapR can allow you to leverage the existing skills of your staff to make these transformations easier.

INTRODUCTION

The journey to modernizing your data analytics strategy isn't always a straight line, and it depends on your starting place. Organizations typically fall into one of these categories:

- Those with little data analytics capacity but a great deal of meaningful data to analyze
- Those who have data analytics capacity but fall short on delivering the expected results and value—these organizations frequently have data in silos throughout the organization
- Those who have moved to a cloud data warehouse or Data Warehouse as a Service model, and are often challenged with ever-growing infrastructure costs

Each type of organization confronts a different set of challenges to reach its data analytics goals. There are myriad options in the business intelligence space and, depending on their maturity level and skills, organizations face a series of decisions around cost and technology to best meet their needs. Let's look at how each type of organization can build the right data analytics solution for their business.

> Vertica integrates with Hadoop via Vertica SQL on Apache Hadoop, which offers the fastest and most enterprise-ready way to perform SQL queries on Hadoop data without moving the data.

Vertica SQL for Apache Hadoop and MapR can work together to meet the data needs of any company. Vertica offers a scale-out, massively parallel processing data warehouse with in-memory query execution, and native ANSI SQL with in-database machine learning functionality. Hadoop provides a converged data platform that can ingest petabytes of data and can support semi-structured and unstructured data types. Vertica integrates with Hadoop via Vertica SQL on Hadoop, which offers the fastest and most enterprise-ready way to perform SQL queries on Hadoop data without moving the data. Together they give companies a powerful combination to perform advanced analytics on massive volumes of data wherever it resides with the SQL skills that they have in their organization.

GREENFIELD ANALYTICS

A firm that's new to this journey has multiple options and minimal technical debt, but its decisions should be driven by its source data and its business goals. An example of such a firm might be a startup Internet-based company with a mobile app that captures data in JSON format. It might also look at data from the web, and ap-plication interactions might be processed in its analytics environment.

This is a firm that likely has decent software development skills but limited experience with building a data model or executing analytical tasks. Depending on the volume of data, it could deploy a Vertica as the core data warehouse and, as its data volumes grow over time, add Hadoop nodes for longer-term storage of its data.

Typically, such a firm would leverage a business intelligence tool like Tableau or QlikView to provide a presentation layer for reports. From a skillset perspective, the firm would want to ensure that it had SQL development expertise and, as it moved into deeper analysis of its data, it should evaluate adding data science skills to move from real-time to predictive and eventually prescriptive analytics.

SILOED DATA—ALL THE HADOOP CLUSTERS

Within large organizations, this challenge typically results when funding for such projects comes from the business side of the organization rather than a central IT function. Each business function declares its Hadoop cluster its own sandbox, and doesn't want to let other parts of the organization share its cluster. This leads to two major problems—proliferation of Hadoop clusters, and data silos throughout the organization. Or, even worse, the data doesn't make it into a Hadoop cluster and lives on file shares or in Excel spreadsheets. In any case, the company isn't maximizing its data insights by centralizing all its data.

By moving to a converged data platform, organizations can lower their total cost of technology ownership, offer a common analytics platform, and gain deeper insights by treating organizational data as a holistic data set, allowing data scientists to see a broad picture across the firm. Many large enterprises have moved to a model of storing all data in a "data lake" using Hadoop, and then adding other analytic tools on top. This gives the advantage of a central data store and allows for granular security. Vertica for SQL on Hadoop provides the performance and functionality that organizations have often lacked when trying to perform analytics on their Hadoop environment.

OVERLOADED WITH DATA CAPACITY—MY CLOUD BILL IS TOO HIGH

Cloud computing provides many benefits, including workload flexibility, operational efficiency, and ease of use, to most organizations. However, a major challenge is predicting monthly costs, especially for Platform-as-a-Service (PaaS) offerings. It can be difficult because you can't simulate your environment in either virtual machines or on-premises hardware. This may not always be a problem, but in the case of rapid growth it can lead to unpredictable costs, which no business wants. In some cases, the platform may simply not be able to meet the performance requirements of the customer without drastically increasing compute spend.

> Rather than constantly adding expensive compute resources as data volumes grow, Vertica gives customers the control and freedom to tune queries and optimize performance.

In an ideal cloud world, customers would increase their workloads in a linear fashion. As the workload grows, they'd add nodes and possibly resize existing nodes. Both Vertica and MapR allow you to scale your cluster as your data volumes and workload grow. Rather than constantly adding expensive compute resources as data volumes grow, Vertica gives customers the control and freedom to

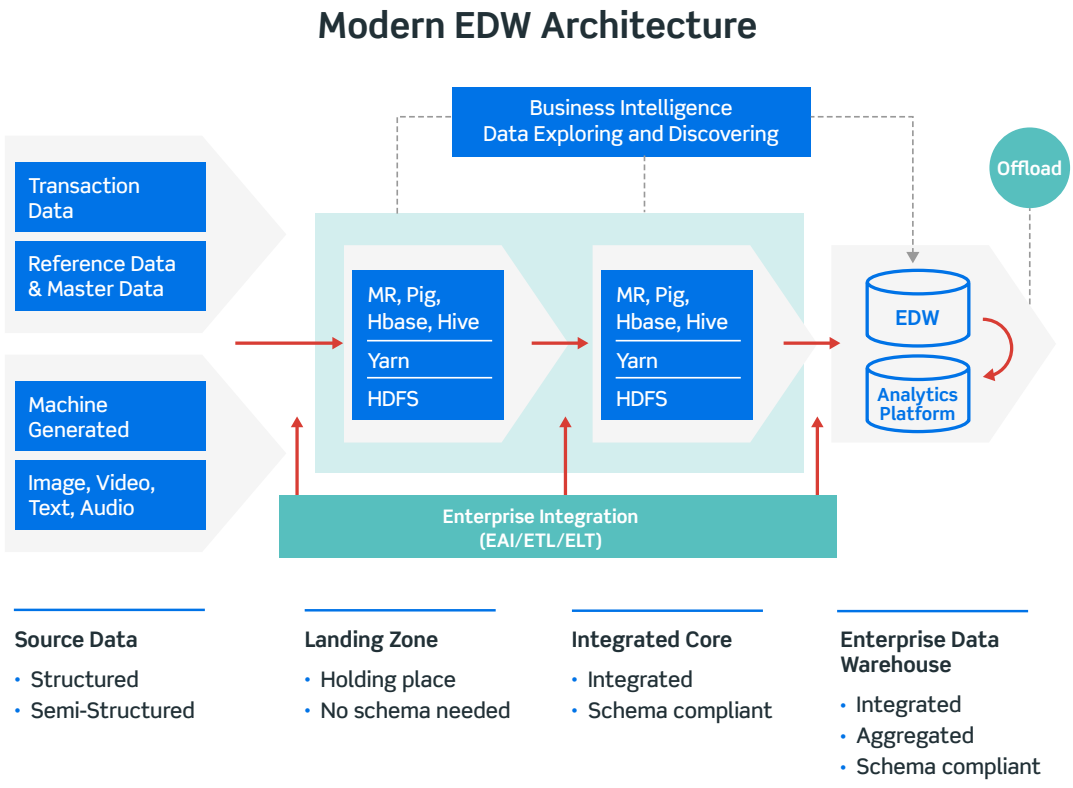

Figure 1: A modern enterprise data warehouse architecture

tune queries and optimize performance. Vertica can typically run on smaller amounts of hardware than other MPP data warehouse platforms. Whether you're on-premises or in the cloud, this lets you predictably manage your costs. If your workload is seasonal, for example, you can easily scale your cluster up or out to meet peak demand.

CHANGING ANALYTICS REQUIREMENTS

A challenge many organizations face is moving from traditional, historical analytics to a more modern, near-real-time analytics system. The major difficulty in making this shift is that it requires changing the extract, transform, and load (ETL) process for the data warehouse from something that runs overnight or every four hours into a process that manages streaming data. This challenge is twofold—the ETL process needs to be rebuilt from the ground up, and the data warehouse needs to able to manage streaming data.

> Traditional data warehouses are designed and optimized for batch loading, not for streaming.

Traditional data warehouses are designed and optimized for batch loading, not for streaming. Vertica provides high performance for ingesting data streams at low latency, whether it's sentiment analysis data from a new social media campaign or sensor data coming from Internet of Things (IoT) devices, using its streaming message bus technology. It also allows for data to be streamed to other targets.

VERTICA AND HADOOP—BETTER TOGETHER

While Hadoop is a powerful platform, its ecosystem is challenging, especially to non-developers. Many of the data interfaces are problematic for business analyst users. And while there are some SQL tools that work with Hadoop, they can be limited and may not run a full set of queries to perform the needed analysis. However, such users are able to query Hadoop tables through Vertica, which allows external tables to be created that can be queried using standard ANSI SQL functionality. This lets you leverage your existing IT and analyst staff to modernize your analytics practice.

PROCUREMENT OPTIONS

Modern IT offers an abundance of hardware and software options. Depending on the nature of your business, you may want to acquire hardware and software through Hewlett Packard Enterprise as a capital expense. This gives you the tax advantage of deprecation, as well as full ownership of your solution.

> Users are able to query Hadoop tables through Vertica, which allows external tables to be created that can be queried using standard ANSI SQL functionality.

Consider HPE Elastic Platform for Analytics (EPA) that enables independent scaling of compute and storage through infrastructure building blocks that are optimized for density and disparate workloads. Combine with HPE MapR converged data platform that offers data services for ingesting, storing, and managing data.

If you prefer a more cloud-based model for meeting your IT needs, choose HPE GreenLake that aligns to your capacity usage for easy scale-up. This allows you to purchase hardware, software, and support on an ongoing basis in a consumption-based model, which offers the ultimate flexibility as you can light up more hardware as your data needs grow.

Vertica is a software solution, completely free from underlying infrastructure, and can be installed and run on the right type of HPE hardware for your organizational and budgetary needs.