pavilion data

Solution Brief

Solution Overview

Modern Infrastructure for Spark-based Applications

Key Benefits

- Improve accuracy of decisions by accessing high volumes of data for analysis
- Mix multiple Spark-based analytics workloads using a common infrastructure
- Increase operational flexibility
- Deliver more application density per rack
- Reduce costs by taking advantage of shared resources and easier management.
- Run multiple modeling operations on the same data set at full speed

Pavilion Overview

- Fastest block storage for flexible private cloud deployments
- Latency of direct-attached SSDs
- Up to 920 TB in 4U
- Frictionless deployment
- Data resiliency & high availability
- Space-efficient instant snapshots and clones
- Thin provisioning
- Pay-As-You Grow scalability
- Expand for capacity or performance, independently
- Increase storage utilization up to 10X or more

We are living in an era of data deluge and as a result, the term "big data" is appearing in many contexts, including meteorology, genomics, complex physics simulations, biological and environmental research, finance, IoT and healthcare.

Apache Spark is an open source cluster computing framework for large-scale data processing. It provides parallel distributed processing, fault tolerance and scalability for big-data workloads.

Storage Challenges and Apache Spark

Many challenges exist related to data management and data storage in large scale data analytics platforms. Some of these challenges may be:

- Inferior storage performance due to the amount of throughput required by parallel, distributed, clustered applications
- Stranded capacity that is held captive in individual servers
- Unpredictable performance due to data management being performed on application servers
- Inability to scale compute, storage, and network resources independently
- Difficulty in providing backup copies of a clustered database, or copies for test/dev usage

It turns out that all of these challenges can be overcome by using the appropriate shared storage solution.

NVMe Storage for Spark-based Big Data Workloads

NVMe is a new storage technology and it is inherently parallel. It is 250 times more parallel than SAS and 2000 times more parallel than SATA. In addition, modern web (transactional) and Machine Learning/AI (real-time analytics) applications are also built upon massively parallel and clustered databases and filesystems because of the performance requirements of these applications.

By leveraging NVMe in Spark-based environments using a storage platform that supports data management operations, organizations can now gain much more value from their data. Multiple applications and analytic workloads can easily be applied to the same data set, or multiple data sets, using common rack-scale infrastructure.

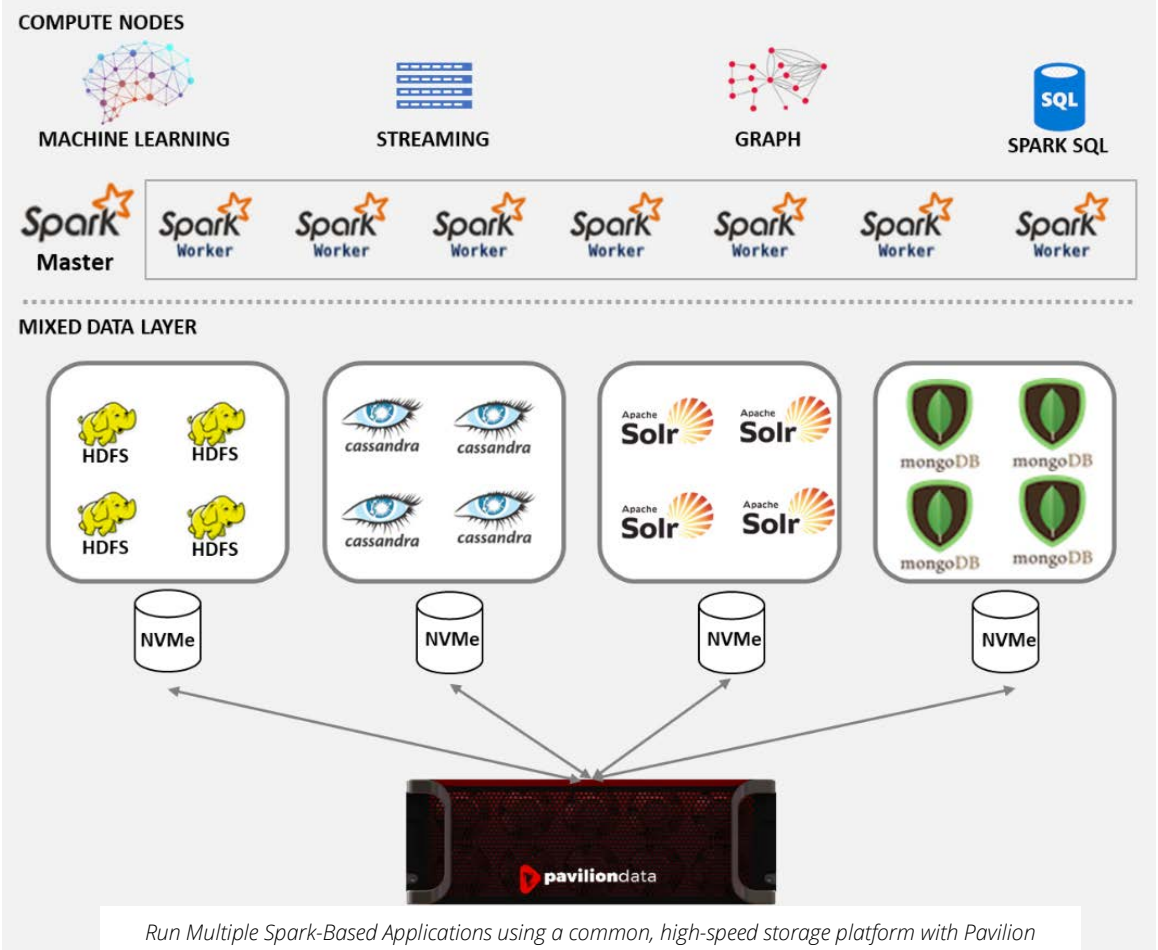
Operational Benefits

- Pavilion instant, full-speed, zero-space Snapshots and Clones can be used to create additional copies of big datasets. This allows multiple workloads to concurrently run on the same version of the data, without degradation or making copies of data over a network
- Reduce management complexity by moving storage to a shared appliance where a low latency pool of shared storage can be centrally managed
- Manage the Pavilion Array through a UI interface, and/or supported by REST API, which allows integration into existing management frameworks

Infrastructure Benefits

- Deploy 'storage-less' 1U servers to deliver more application compute density per rack
- Save cost in several areas including hardware acquisition, rack space, power and cooling
- Pavilion's Thin Provisioning support allows for less raw flash capacity to be installed
- Scale compute, storage and network resources independently to meet diverse application requirements

The diagram represents the reference architecture for a modern Apache Spark implementation across complex persistent layers for modern applications, leveraging scalable NVMe-based shared storage resources.



Run Multiple Spark-Based Applications using a common, high-speed storage platform with Pavilion

pavilion data

Pavilion Data Systems, Inc.
2550 N First St., Suite 250, San Jose, CA 95131
E-mail: sales@paviliondata.com