Guide to Supporting On-Premise Spark Deployments with a

Cloud-Scale Data Platform

Apache® Spark™ has become one of the most rapidly adopted open source platforms in history. Demand is predicted to grow at a compound annual rate of 67% per year between 2017 and 2020, with the cumulative Spark market valued at more than $9 billion during that period, according to the research firm MarketAnalysis.com.[1]

There are several reasons for this staggering growth. One is that Spark is "a damn good technology," according to MarketAnalysis.com, providing users with a fast in-memory data engine that supports a wide range of expressive development APIs. Other key factors driving the growth of Spark include:

- **The rising importance of big data analytics** in general and the specific preeminence of Hadoop® as an analytics platform.
- **Spark's ability to speed analytics applications** by orders of magnitude, shaped by characteristics such as versatility and ease of use.
- **Spark's flexibility** in allowing developers to write applications in Java®, Scala, Python®, or R.
- **The maturation of Spark technology** at a time when billions of Internet of Things (IoT) devices are about to hit the market.

When a technology grows this quickly, however, it can often create challenges for the IT team. Spark is no exception. To date, a large portion of Spark deployments have taken place on public cloud platforms, predominantly Amazon Web Services (AWS)®.

1 "Apache Spark Market Forecast, 2017-2020," MarketAnalysis.com, Feb. 11, 2016

PURESTORAGE® Hello Possible™

TechTarget® | Custom Media

There are several reasons for the use of public cloud, including ease of use and the ability of development teams to quickly acquire the infrastructure resources they require. In addition, AWS has done a good job of providing tools to help software teams be efficient in using Spark as a development platform and as a processing engine for big data analytics.

The Risks of Public Cloud

Over the longer term, however, relying completely on the public cloud for Spark can be a risky proposition for IT teams and for the various users they support, including data scientists, engineers, DevOps team members, and others. There are several potential challenges in using public cloud, which many IT teams are already beginning to confront, including:

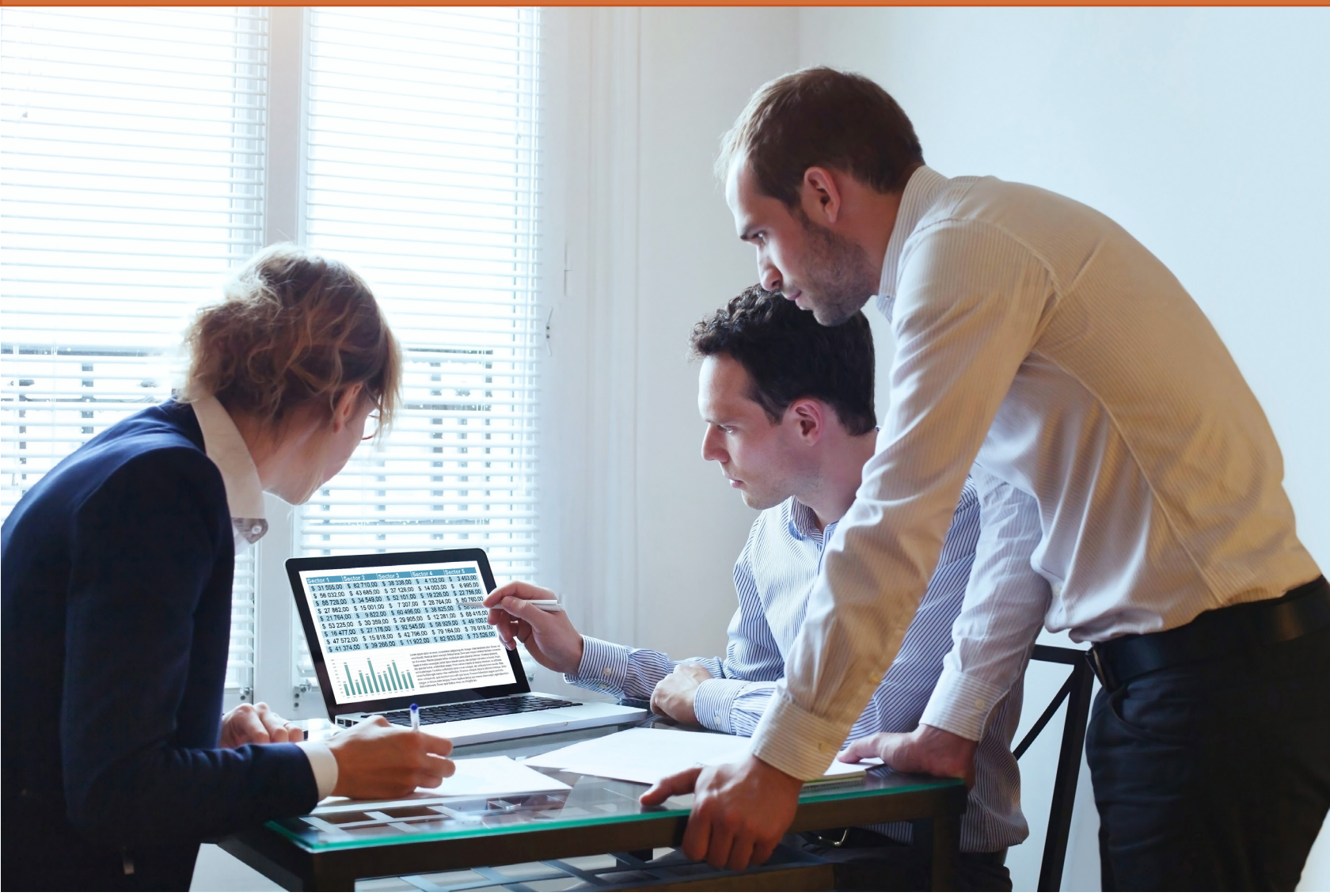
- **Costs:** As the deployment gets larger the costs can grow significantly. And, while AWS cuts prices on a regular basis, the reality is that the costs involved in hosting Spark in the public cloud can become prohibitive. Some companies are already spending millions of dollars every month on public cloud services just to support Spark. Bringing Spark in-house can be a more cost-efficient solution, provided you have the right technology in place.
- **Control:** This is always an issue for IT departments, and it is becoming more of a challenge in the cloud era. When you rely on a public cloud provider, you are dependent on that provider to deliver everything you need, whether that is capacity, performance, or security. The individuals using Spark tend to be major contributors to your organization—among the most important drivers of revenue and innovation. Do you really want to have an outside provider exert that much control over such a core aspect of your business?
- **Performance:** This is another factor that can fall under the area of control, in the sense that the IT team would like to be able to control the amount of performance delivered to Spark users and not be reliant on a public cloud provider to deliver the speed and IOPS required for specific workloads. There will be some workloads that require higher levels of performance than what is available through the public cloud, particularly when you consider issues around bandwidth and connectivity to pubic cloud resources. If the IT team is supporting Spark through on-premises infrastructure, it has a better opportunity to deliver the right performance to the right workload at the right time.

Storage Challenges in Supporting Spark On Premises

One of the biggest challenges in bringing Spark on premises has been the lack of a viable alternative, particularly when it comes to the storage infrastructure required to support this type of fast in-memory data engine. Most IT teams looking to bring Spark in-house would typically start with the Hadoop Distributed File System (HDFS) on direct-attached storage (DAS), which has been the de facto storage solution for Spark.

But DAS is not an ideal solution by any means. First, it is extremely expensive, and for a technology growing as quickly as Spark, scaling DAS can be prohibitive. In addition, DAS is not efficient or flexible because it couples compute and storage as the unit of scaling. If you have to increase storage capacity, you must buy more compute resources, even if you don't need them. That impacts costs as well as IT efficiency. Finally, using DAS can have a negative impact on manageability and resiliency, depending upon how IT manages functions such as data services.

So, while many IT leaders may have been inclined to move Spark out of the public cloud and into an on-premises data center, they haven't been able to pull the trigger because the in-house storage technology alternatives haven't quite been up to the task.



A Next-Generation Storage Technology to Support Spark

The answer to successfully bringing Spark deployments in house, at least from the standpoint of storage technology, is a solution that can deliver on the following key characteristics:

- **Uncompromised performance:** Extreme and consistent performance has to be delivered concurrently across potentially large data sets, which can reach multi-petabyte levels. In order to fully leverage the benefits of Spark, the storage infrastructure must not only deliver high capacity, but it must also enable extreme concurrent performance in terms of bandwidth, IOPS, low latency, and high availability.
- **Simple/independent scalability:** You want to be able to add storage capacity when you need it, without also having to pay for incremental compute capacity that you may not need. This is in sharp contrast to moving Spark on premises using DAS solutions.
- **A shared storage model:** One of the biggest challenges with Spark is the sheer volume of storage capacity required, particularly with the growth of unstructured data and the IoT. As IT leaders know too well, data is expected to keep growing at a blazing pace. Team members in typical Spark use cases—whether they are scientists, software developers, engineers, or others—need to share information and work in a highly collaborative environment. Shared storage enables concurrent, independent access to common datasets.

- **Cost efficiencies:** Cost is tied directly to IT control. If you are using a public cloud service such as AWS, you are beholden to that provider for your cost structure. However, if your goal is to bring Spark on premises, the cost structure has to be at least competitive with AWS', if not less expensive. Therefore, you need a shared storage solution that not only delivers the right amount of performance and scale, but is also cost efficient.
- **Management simplicity:** One of the reasons many organizations have turned to the public cloud is ease of use. Just as an on-premises solution must be competitive on cost, it must also deliver simplicity across the board—in the initial deployment, scaling, ongoing support, maintenance, and upgrades. You should be able to upgrade without having to rip and replace your existing infrastructure, and you should be able to adopt new technologies without having to go through lengthy and risky data migrations.

Innovative Technologies to Address Spark Storage Challenges

The combination of extreme concurrent performance, capacity, scalability, and simplicity has been a challenge to storage vendors. Attempts at building solutions—including block-based all-flash arrays, parallel file systems, and hybrid architectures—have typically fallen short in one or more key areas.

With the growth of unstructured data and the challenges of modern workloads such as Spark, IT teams have seen a clear need during the past few years for a new type of all-flash storage solution, one that has been designed specifically for users requiring high levels of performance in file- and object-based environments.

Pure Storage®, one of the leading innovators in block-based all-flash arrays, is among the first vendors to develop a solution that meets the extreme performance challenges of file- and object-based environments. With FlashBlade™, Pure has created a new paradigm that combines extreme performance, simple manageability, and high density for the most demanding workloads.

FlashBlade deploys an innovative scale-out architectural model for all-flash arrays. The architecture is built on three core components: the blade, elasticity software, and an elastic fabric. A single FlashBlade chassis can support up to 15 blades, and each blade can be configured with either 8.8 TB or 52.8 TB of raw flash storage, supporting tens of billions of files and objects in an extremely small all-flash footprint.

FlashBlade addresses performance challenges in Spark environments by delivering the consistent performance of all-flash storage with no caching or tiering, as well as fast metadata operations and instant metadata queries. Each FlashBlade chassis supports up to 15 GBps of bandwidth and 500K NFS operations per second. It was designed so that anyone can install it.

One of the other advantages of FlashBlade for on-premises Spark deployments is its simple scalability. Every dimension of the system can scale linearly with the system as it grows. This includes IOPS performance, bandwidth, metadata performance, NVRAM, and client connections. FlashBlade is also inherently multiprotocol, giving users more flexibility to use legacy file access protocols, as well as newer object protocols and application-specific protocols.

In addition to performance, capacity, simplicity, and scalability, FlashBlade delivers cost efficiencies, particularly for on-premises Spark deployments. It takes up to 20X less space and uses up to 10x less energy than traditional storage systems. Companies can also take advantage of Pure's Evergreen™ Storage model to reduce costs and use an Opex versus Capex model. With Evergreen Storage, organizations can avoid rip-and-place upgrades and data migrations, while always remaining current with the latest technology.

Taking the Next Step: Leveraging the Business Benefits of Spark

For users in Spark environments and the IT teams supporting them, FlashBlade provides an opportunity for a new deployment model that brings Spark on premises. This delivers a wide range of benefits: IT teams have much greater control over costs, security, and performance; and strategic teams and business units—such as DevOps or scientific research—can work collaboratively using a shared storage platform.

By leveraging the consistent and extreme performance of Spark, development teams, engineers, data scientists, and others can be far more creative and innovative. They can conduct more iterations and tests in shorter time frames, using larger datasets. Further, they can leverage Hadoop, big data analytics, and the IoT to develop new business services and drive profitability.

In an era when speed to market has become an important competitive differentiator, these factors can have a far-reaching positive impact on the business. By improving performance and simplicity, the IT department will also help to create a more exciting and compelling work environment for users who are important contributors and innovators for the business.

Conclusion

Apache Spark has become a critical tool for all types of businesses across all industries. It is enabling organizations to leverage the power of analytics to drive innovation and create new business models.

The availability of public cloud services, particularly Amazon Web Services, has been an important factor in fueling the growth of Spark. However, IT organizations and Spark users are beginning to run up against limitations in relying on the public cloud—namely control, cost and performance.

Until now, there hasn't been a great alternative to the public cloud because the storage infrastructure has been a gating factor in meeting the challenges of Spark, particularly in performance, scalability, and ease of use. The paradigm has shifted, however, with the introduction of FlashBlade from Pure Storage.

FlashBlade represents a new scale-out architectural model for file- and object-based storage environments, delivering previously unattainable levels of concurrent performance in a shared storage platform that is cost-efficient; high capacity; and easy to deploy, scale, and manage.

With FlashBlade, IT teams can take control over their Spark environments and move them on premises. Spark users can be more creative and productive, and the overall business can benefit from increased innovation and accelerated speed to market.

For more information on how your organization can leverage Pure Storage FlashBlade technology to enable and support an on-premises Spark deployment, please visit Pure Storage at PureStorage.com/Analytics.
