ELASTIC DATA PLATFORM

A scalable and efficient approach to provisioning analytics sandboxes with a data lake

ESSENTIALS

- **Powerful**: provide read-only data to anyone in the enterprise while avoiding data sprawl
- **Tailored**: create secure, isolated analytics work environments running on the infrastructure of your choice
- **Elastic**: scale up/down while allowing storage to scale independently of compute resources
- **Cloud Ready**: deploy both on-premises and hybrid solutions
- **Scalable**: go from a handful of bare metal servers to rack-scale hyper-converged infrastructure
- **Tool Agnostic**: use the analytics tools that work best for you

Overview

It is common for Big Data deployments to start out on a small scale with the software components installed on bare metal servers with direct-attached storage. As a data lake—part of a system that stores and processes large amounts of data—grows and the organization begins to use it for analytics initiatives, it tends to become disk/storage constrained with little CPU/memory constraints. As organizations expand those deployments beyond 30-40 nodes/servers, the management overhead of those deployments becomes more costly and less efficient. It is also common to deploy multiple Hadoop® clusters to solve this problem by spreading the data around, but with the rapid growth of data, this also becomes inefficient to maintain, and even more so if copies of data reside on multiple clusters.

The cost to organizations is two-fold: buying high-end servers to keep up with the storage demands costs too much, and the time-consumption of maintaining all of those servers running bare metal reduces efficiency. Many organizations have started to encounter this situation. Dell EMC Consulting has designed and proven a cost-effective and scalable solution.

Dell EMC Consulting proposes an **Elastic Data Platform**, with containerized compute nodes, decoupled storage, and automated provisioning. Additionally, as clusters proliferate and the number of differing data applications (i.e., Hadoop, Spark, data science, and machine learning platforms) increases, enforcing access restrictions and policies becomes critical when scaling the environment.

The time-consuming nature of hand-building a new environment for each user (acquire compute node with storage, install operating system, install Hadoop version, install applications, patch, test and deploy, and then securing all of those components) can compound the chances of errors. Whereas, moving to a containerized, compute-and-storage configuration with a front-end management system, and deployment of a centralized policy enforcement engine can reduce the time to deploy a new Hadoop cluster from days or weeks to hours in a repeatable manner.

Platform Flexibility

Dell EMC Services understands that many organizations have existing and maturing Big Data deployments. These deployments contain infrastructure, applications and carefully considered customizations. It has become increasingly difficult for enterprises to keep up with the pace of change in Big Data.
It’s a rapidly evolving landscape for both data science and IT teams—with a steady stream of new products, new versions and new options for frameworks like Hadoop and Spark. Data scientists and developers want flexibility and choice, with on-demand access to these new Big Data technologies. Big Data architects and IT managers are under pressure to support these new innovations and the ever-changing menagerie of tools, while also providing enterprise-grade IT security and control. Rather than ripping and replacing these investments, Dell EMC recommends augmenting these deployments with enhancements that provide greater user access, elastic scalability and strong security/compliance.

Solution

The solution deploys Docker containers to provide the compute power needed along with an Isilon storage cluster using HDFS. The solution will use software from BlueData, which provides the ability to spin up instant clusters for Hadoop, Spark, and other Big Data tools running in Docker containers. This enables users to quickly create new containerized compute nodes using predefined templates, and then access their data via HDFS on the Isilon system. With a containerized compute environment, users can quickly and easily provision new Big Data systems or additional compute nodes to existing systems—limited only by the availability of physical resources. By consolidating the storage requirement onto an Isilon storage cluster, the need to have redundant storage reduces the data replication factor of 200 percent to only 20 percent overhead. It also enables the sharing of data between systems and enterprise level features for the data such as: snapshots, disaster recovery replication, and automated data tiers to move data to appropriate storage tiers as the data ages.

After implementing a containerized Big Data environment with BlueData, Dell EMC Consulting deploys a centralized policy engine (provided by BlueTalon). Then, we create and deploy a common set of policies about who can access what data, and how via simple rules (e.g., allow, deny or mask the results) via enforcement points to all of the applications accessing the data. This ensures the definition and enforcement of a consistent set of rules across all data platforms, ensuring data governance and compliance, by only allowing users or applications access to the data to which they are entitled.

The resulting solution is a secured, easy-to-use and elastic platform for Big Data with a flexible compute layer and consolidated storage layer that deliver performance, management, and cost efficiencies unattainable using traditional big data architectures.

Principles

Building on the success of existing data lake deployments, organizations will provide users and departments with a variety of Big Data and data science workloads to exploit the data to perform traditional query and advanced analytics. There are five key principles used to guide the enhancements:

• **Easy Data Provisioning**
  Provide read-only access and scratch pad data to anyone within the organization while preventing data sprawl and duplication.

• **Tailored Work Environments**
  Isolate environments between users to ensure data integrity and reliable compute performance tailored with a variety of tools for many different workloads—assuring quality of service.

• **Scalability**
  Ensure compute environment performs elastically and scales horizontally to meet business demands and deliver high quality of service.

• **Data Security**
  Enhance security, governance and access controls while maintaining ease of use.

• **Cloud Ready**
  Establish on-premises model while preparing for hybrid on/off-premises solution.
Solution Details

Separating Compute and Storage

Although decoupled storage is not required with the Elastic Data Platform, once the data set becomes larger than a few hundred terabytes, Dell EMC’s Isilon solution offers a compelling ROI and ease of use with scalability. Isilon provides several capabilities that extend the value of the Elastic Data Platform:

- Separation of the storage allows for independent scaling from the compute environment
- Native erasure encoding reduces the overhead from 200 percent to 20 percent
- Creation of read-only snapshots, on some or the entire data set, effectively requires no additional storage space
- Auto-tiering of data (i.e. Hot, Warm and Cold) maximizes performance/throughput and cost effectiveness

Deployment, Orchestration, and Automation

When deploying a cluster, BlueData quickly spins up compute clusters while Isilon HDFS provides the underlying storage for the compute clusters. Clusters can be deployed using various profiles that are based on the end user’s requirements (e.g., a cluster could have high-compute resources with large memory and CPU and have an average throughput storage requirement).

Decoupling storage and isolating compute provides the organization with an efficient and cost-effective way to scale the solution (i.e. scaling compute and storage resources independently), providing dedicated environments suited for the various users and workloads coming from the business.

Tenants within BlueData are logical groupings defined by the organization (i.e., different departments, different business units, different data science and analyst teams) that have dedicated resources (CPU, memory, etc.), and can then be allocated to containerized clusters. Clusters also have their own set of dedicated resources (coming from the tenant resource pool).

Applications that are containerized via Docker can be made part of the BlueData App store and can be customized by the organization. Those application images are made available to deploy as clusters with various “Flavors” (i.e., different configurations of compute, memory and storage.)

The data residing on HDFS is partitioned based on rules and definable policies. The physical deployment of the Isilon storage allows for tiering, and the placement of the data blocks on the physical storage is maintained by definable policies in Isilon to optimize performance, scale and cost.
Isilon is configured to generate read-only snapshots of directories within HDFS based on definable policies.

Users gain access to the data through the DataTap functionality in BlueData’s software. A DataTap can be configured to tap into any data source; each DataTap is associated with Tenants and can be mapped (aka “mounted”) to directories in Isilon. These DataTaps can be specified as Read-only or Read/Write. DataTaps is configured for connection to both the Isilon Snapshots and writeable scratch-pad space.

Once users have finished their work (based on informing the administrators that they are finished, or based on their environment time being up), the system removes temporary space in Isilon and removes or reduces the size of the compute environment so that those resources can be made available to other users.

Centralized Policy Enforcement

The difficulty many organizations face with multiple users who access multiple environments, with multiple tools and data systems, is the consistent creation and enforcement of data access policies. Often, these systems have different inconsistent authorization methods. For example, a Hadoop cluster may be Kerberized, but the MongoDB cluster may not be, and the Google BigQuery engine would have its own internal engine. This means that administrators must create policies for each data platform and independently update them every time there is a change. In addition, if there are multiple Hadoop clusters and/or distributions, then administrator must define and manage the data access independently and with capability inconsistency across the system.

The solution to this is to leverage a centralized policy creation and enforcement engine, such as BlueTalon. In this engine, the administrator simply creates the policies once by defining the access rules (i.e., Allow, Deny and Mask) for each of the different roles and attributes for the users accessing the system. Then, distributed enforcement points are deployed to each of the data systems that read the policies from the centralized policy engine, and enforce them against the data. This greatly simplifies the overall Big Data environment and allows for greater scalability while maintaining governance and compliance without impacting user experience or performance.

Integration

Alone, any of the above components will provide value to the organization. However, to truly achieve the goals of the 5 Principals, the solution must provide a level of automation of the components and integration into the existing enterprise Big Data environment. Dell EMC Consulting automates analytical sandbox creation, data provisioning via read-only snapshots and wraps security policies around those environments. Further, an open and extensible interface is available for integrating with existing Big Data systems within the enterprise to enable self-service by end users.

A typical enterprise will have existing IT self-service capabilities through ticketing systems or Portals (i.e. ServiceNow.) Additionally, ingestion, processing, and meta-data management systems are often in place to discover, move and track the data. Many organizations also require the automation of Kerberos certificate generation for any Hadoop cluster as well as the registration of those clusters in the corporate DNS. The Elastic Data Platform provides an interface to integrate into these systems to enhance the overall capabilities of the organization’s Big Data solution.

When a user requests a new Big Data environment, The Elastic Data Platform will automatically provision that environment based on the parameters specified in the request. Then, that environment will be connected to two data stores—one that provides a read-only view against the Hadoop data set, and another that provides writable space for sandboxing. Finally, the entire environment is secured and the data access is restricted based on policies automatically applied to the environment. The seamless experience for the end user has streamlined a process that takes weeks to months in the traditional enterprise down to hours. This provides greater productivity for the Data Scientists and Analysts who will in turn deliver greater value back to the business because they will spend time performing work rather than waiting for environments to be provisioned.
Summary

The Dell EMC Elastic Data Platform is a powerful and flexible approach to help organizations get the most out of their existing Big Data investments. Its scalability, elasticity and compliance support the ever-growing needs of the business. It provides fast and easy provisioning, simplified deployments, cost sensitivity and assurance that governance and compliance requirements are being met.