 SERVICE OVERVIEW

ELASTIC DATA PLATFORM
Provision and optimize analytics environments at speed and scale

ESSENTIALS

• Offers a powerful and flexible approach to managing and optimizing Big Data infrastructure for exploratory analytics and flexible workloads

• Combines industry-leading Dell EMC infrastructure, Consulting services and partner software for an integrated, end-to-end solution that augments existing Big Data investments

• Enables users to work with the tools of their choice and the data they require through a self-service delivery model

Overview

Big Data deployments often begin on a small scale with bare metal servers and direct attached storage. As these systems grow, organizations often struggle to keep pace with business demand, due to limited resources. Lagging behind and delays can be costly, especially if it means data scientists are waiting for an environment to be provisioned rather than discovering untapped value in the data.

Dell EMC offers an Elastic Data Platform, which provides self-service analytics in a cloud-like environment for exploratory analytics and flexible workloads. The solution reduces costs by optimizing the physical infrastructure, frees IT resources up from tedious manual processes, and enables the business to get to actionable insights faster with provisioning of the data and environments they need in minutes.

Business Challenge

In today’s digital age, the Big Data landscape is rapidly evolving for both data science and IT teams—with a steady stream of new products, tools and frameworks. Data scientists and developers want flexibility and choice, with on-demand access to new Big Data technologies such as machine learning and artificial intelligence. 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.

Under these conditions, it has become increasingly difficult for enterprises to keep up with the pace of change in Big Data. Bare metal servers with direct-attached storage and the data lake can quickly become disk/storage-constrained as an organization’s use of the data expands. As nodes/servers are added, the management overhead becomes costly and inefficient, not to mention the costs of the servers themselves.

A common approach to address this problem is to spread the data around across multiple Hadoop® clusters. However, with the rapid growth of data, this also becomes inefficient to maintain, and even more so if copies of data reside on multiple clusters. As clusters proliferate and the number of analytics and data science applications and tools increases, enforcing access restrictions and policies can also be challenging as the environment scales.

Additionally, the time-consuming nature of manually building a new environment for each user—to acquire a compute node with storage, install the operating system, install the Hadoop version and applications, patch, test and deploy, and then secure all of those components—can compound the chances of errors and cause costly delays to the business.
Solution

The Elastic Data Platform is a proven and cost-effective solution that enables organizations to address these challenges at speed and scale. It is designed to extend and augment an organization’s existing Big Data investments with workload-specific infrastructure, intelligent software and end-to-end automation—no ‘rip and replace’ required. The integrated solution combines industry-leading Dell EMC infrastructure and partner software, delivered with the strategic guidance and expert execution of Dell EMC Consulting to accelerate the time to value.

The Elastic Data Platform is built around 5 key principles:

- **Ease of data provisioning**
  Read-only access and scratch pad without data duplication

- **Scalability**
  Elastic compute environment to scale based on business demand

- **Tailored work environments**
  Isolated sandboxes and tool flexibility

- **Data security**
  Security, governance, and access controls while maintaining ease of use

- **Cloud-ready**
  On-premise or hybrid cloud operability

![Figure 1. Elastic Data Platform Principles](image)

The Elastic Data Platform deploys Docker containers to provide the compute power needed along with scale-out Network Attached Storage via the Dell EMC Isilon storage cluster. The solution uses software from BlueData, which provides the ability to spin up clusters, nearly instantly, for Hadoop, Spark, AI and machine learning libraries, and other Big Data tools running in Docker containers. This enables users to quickly create new analytical environments using predefined templates and then access their data via Hadoop 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 on to 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, and automated data tiers (i.e., hot, warm, and cold) to move data to appropriate storage tiers as the data ages.

Using BlueTalon as the centralized policy engine, a common set of policies is developed and deployed defining who can access what data and how, via simple rules and enforcement points for 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 delivers performance, management and cost efficiencies unattainable using traditional big data architectures.

Solution Details

Separating Compute and Storage

The massively parallel nature of AI and machine learning workloads’ high data throughput and concurrency rates can create I/O bottlenecks and starve GPU’s of data. While 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 capabilities that extend the value of the Elastic Data Platform:

- Separation of the storage allows for independent scaling from the compute environment
- Auto-tiering of data (hot, warm and cold) maximizes performance/throughput and cost effectiveness
- Maximized data throughput as well as GPU utilization to prevent GPU starvation
- Faster training and validation of models
- Higher model accuracy with larger and more detailed data sets
Deployment, Orchestration, and Automation

When deploying a cluster, BlueData quickly spins up compute while Isilon provides the underlying storage for the compute clusters. Clusters can be deployed using various profiles that are based on the end user’s requirements.

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

Tenants within BlueData are logical groupings defined by the organization (e.g., different departments, different business units, different data science and analyst teams) that have dedicated resources like GPUs and memory, 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,” such as different configurations of memory and storage.

The data residing in Hadoop 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 to optimize performance, scale and cost.

Isilon is configured to generate read-only snapshots of directories within Hadoop.

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, isolated, and secured with Tenants and can be mapped 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. 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 the 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 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 and without impacting user experience or performance.

Alone, any of the above components will provide value to the organization. However, to truly achieve the goals of the five design principals, the solution must provide a level of automation of the components and integration into the existing enterprise Big Data environment. The Elastic Data Platform 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 (e.g., 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 minutes. 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 optimize their environments for exploratory analytics and flexible workloads. Its scalability, elasticity, and compliance support the ever-growing needs of the business. Additionally it provides fast and easy provisioning, simplified deployments, cost efficiencies, and automated processing.

To accelerate the time to value, our experienced consultants will work with your organization to define the solution architecture and then integrate the Elastic Data Platform and Big Data-as-a-Service capabilities into your Big Data environment. The platform can be stood up in six weeks with our accelerator option and fully operational in three to six months, enabling higher data scientist productivity and increased opportunities for data monetization at speed. Dell EMC also offers Ready Solutions for Big Data, which includes the accelerated implementation option, along with Dell EMC servers, networking and support services.

Are you ready to harness the power of big data and analytics to transform your organization? Dell EMC offers a comprehensive portfolio of Big Data Consulting services from strategy through to implementation and ongoing support. Contact us to learn more.