Multi-tenant Big Data analytics on demand: One click deploys Hadoop clusters in minutes

Business intelligence and data analytics are now the single biggest area of IT spending for enterprises. Many organizations today are looking to leverage their ever-growing Big Data assets to uncover new opportunities to accelerate their business and gain a competitive advantage, and Big Data technologies are capturing an ever-increasing share of this market, with Hadoop featuring prominently.

Hadoop is an open-source project platform designed to process petabytes of unstructured data using a distributed computational paradigm called MapReduce, in conjunction with the Hadoop Distributed File System (HDFS), Hive, HBase, and several other Hadoop sub-projects. Hadoop’s abilities to handle extremely large, unstructured data sets efficiently and affordably make it a valuable toolkit for enterprises across a number of applications and fields.

But Hadoop has traditionally created operational challenges for organizations, including difficult deployment, poor utilization of storage and/or CPU, inefficient data staging and loading, and a lack of multi-tenancy capabilities. That inability to create performance or security isolation between different users and tenants, as well as a lack of resource containment for different service levels, has stalled the proliferation of Hadoop in multi-tenant private, hybrid and public clouds.

Add to that the technology challenges, such as an inability to run multiple versions on clusters, inadequate security management and the lack of backup and disaster recovery, and service providers are challenged to take advantage of the significant opportunity Hadoop provides as a cloud service offering to their customers.

Figure 1. Classic Hadoop nodes require rigid physical hardware and local storage—a cloud implementation is needed for service providers
Analysts expect about 15% of IT spending to move to the public cloud by 2015. By conservative estimates, the total addressable market for HDaaS in 2012 was $130 million, with growth projected to remain strong in subsequent years.

HADOOP AS A SERVICE
Hadoop-as-a-Service (HDaaS) offers enterprises a cost-effective approach to meeting their big data analytics needs. Service providers (SPs) with existing managed services are ideally positioned to add HDaaS to their offerings portfolio with a convenient yet sophisticated platform that addresses their customers' need for on-demand high-performance unstructured data analytics, while addressing their own challenges associated with managing a complex, dynamic Hadoop environment.

THE EMC HDAA S SOLUTION
The EMC HDaaS solution is a platform-as-a-service solution for technical data scientist end users. It provides a multi-tenant, self-service portal that leverages Pivotal HD, EMC Isilon for shared Hadoop storage and VMware vSphere Big Data Extensions for virtualized Hadoop technologies.

The EMC HDaaS solution lets private and public service providers offer Hadoop-as-a-Service by addressing the challenges SPs face when building an in-the-cloud solution using classic Hadoop.

Classic Hadoop implementations
- Can be hard to deploy and to operate
- Use compute and storage inefficiently
- Are inefficient at data staging and loading
- Lack multi-tenancy support, backup and disaster recovery
- Exhibit cluster sprawl and poor manageability

The EMC HDaaS solution can instantly provision as much or as little capacity as users need to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics research. The EMC HDaaS solution allows data scientists to focus on crunching or analyzing data without having to worry about the time consuming setup, management/tuning of Hadoop clusters and their computational capacity. The solution maximizes the value of SP infrastructure by using virtualization and shared storage for a Hadoop service, rather than rigidly and inefficiently dedicating physical infrastructure.

The EMC HDaaS solution is comprised of a front-end portal that lets users load data into a shared Isilon HDFS storage back end and submit Hadoop jobs and job flows. The HDaaS solution takes care of spinning up a virtualized Pivotal (Greenplum) HD Hadoop cluster, i.e. virtual machines.

PIVOTAL HD
Pivotal HD is EMC’s 24x7 supported distribution of Hadoop that is proven at scale and provides pluggable storage options. Pivotal HD provides a complete data analytics platform including consulting services, training, and global support. Joining EMC Pivotal HD analytics software and a purpose-built data computing appliance enables data scientists to consume your Hadoop solution quickly and easily.

EMC ISILON
Isilon is the first and only enterprise-ready, scale-out NAS storage platform that natively integrates the HDFS protocol.

EMC Isilon was designed to
- Deploy faster with native Hadoop integration
- Support multiple applications and workflows
- Reduce risk with a highly resilient architecture

The EMC Isilon storage used in this solution marries the simplicity of Isilon scale-out NAS storage with the leading-edge analytics tools of Pivotal HD, resulting in a highly integrated, one-stop Hadoop solution that allows data scientists to quickly extract new insight from their data.
A PLATFORM SOLUTION

EMC HDaaS was designed to be a platform-as-a-service solution for data scientists, rather than an end-to-end software-as-a-service analytics solution for end users who want to run certain queries and reports. However, data scientists can leverage the HDaaS platform to build such higher-level services for enterprise analytics consumers.

Verticals that can benefit from the application of Hadoop technology platform in the cloud include financial services, internet retailers, pharma/drug discovery, mobile/telecom, scientific research, and social media.

The solution architecture shown in Fig. 2 consists of three tiers: The users (SP admins, tenant admins, data scientists) all interact with the HDaaS GUI, which is the visible part of the HDaaS portal functionality. Servers are configured in a management resource pool and a cluster resource pool, where the cluster resource pool provides the compute nodes for the Hadoop clusters and can be further subdivided to reflect different service levels (gold, silver, bronze). The management pool is used for the management tools such as vCenter or vCenter Operations Manager, as well as the HDaaS portal.

The integration point of the HDaaS functionality is the HDaaS API. This API exposes all the functionality of the solution and is invoked by the HDaaS GUI, but can also be leveraged by any other user interface front-end.

The HDaaS API in turn uses the APIs of the backend systems to implement its functionality. These backend systems are Isilon, vSphere and vCenter.
Virtualized Hadoop with local storage provides unified operations with shared resources for higher utilization. In addition, elastic resources enable faster provisioning of services.

Virtualized Hadoop with shared storage in Fig. 5, leverages a virtual infrastructure with physical hardware. Built upon a multi-application scale-out storage platform that supports independent scaling, SP’s can add more storage capacity without adding another server (and vice versa).

With EMC Isilon, there is also the added benefit of linear increases in performance as the scale increases. As a result, replication overhead is reduced from 200% to 20%!

CONCLUSION

Geared towards highly parallel processing across a potentially large number of nodes, Hadoop has followed a high-performance computing model in its design.

Service providers who already offer cloud-based services or traditional hosting services are ideally positioned to round out their as-a-service offerings by providing the EMC HDaaS solution to their installed base.

The EMC Hadoop-as-a-Service solution

- Automates one-click, elastic provisioning of multiple, high-performance Big Data analytics environments
- Virtualizes Hadoop infrastructure to provide multi-tenant cloud features, and more efficient use and scaling of decoupled storage and compute resources
- Provides the first and only enterprise-ready, scale-out network attached storage protocol to natively support HDFS
- Simplifies Hadoop cluster management and eliminates unnecessary data loading
- Can improve CPU utilization 5-10 times, and reduce disk space overhead from 200 percent to 20 percent
- Lets data scientists focus on analytics, instead of having to manage complex Hadoop infrastructure
- Lets service providers focus on maximizing resource utilization, minimizing management and per-subscriber costs, while increasing Big Data analytics service consumption

EMC offerings in backup and recovery, enterprise content management, unified storage, big data, enterprise storage, data federation, archiving, security, and deduplication help customers move to and build IT trust in their next generation of information management, enabling them to offer IT-as-a-Service as part of their journey to cloud computing.

LEARN MORE

To find out how EMC products, services, and proven solutions can help solve your business and IT challenges, contact your local representative, authorized reseller, or visit us at www.EMC.com.