Hadoop in the Hybrid Cloud

Presented by Hortonworks and Microsoft
Introduction

An increasing number of enterprises are either currently using or are planning to use cloud deployment models to expand their IT infrastructure. Big data with Apache Hadoop™ is one of the technologies that enterprises are planning to migrate to a cloud model. A hybrid cloud deployment model provides the best of both on-premises and public cloud models and offers benefits and advantages for Apache Hadoop™ in particular.

This paper provides an introduction to hybrid clouds and outlines the important requirements that enterprises need to consider as they migrate their big data architectures to a hybrid cloud deployment model.

These requirements are addressed by the Microsoft and Hortonworks solution, Hadoop in the Hybrid Cloud. The benefits and advantages of the Hadoop in the Hybrid Cloud solution are outlined so that enterprises understand what tools are available to help migrate to a hybrid cloud quickly and efficiently.

Business Trends and Drivers: Cloud Migration

Business and technology are moving at an extremely fast pace today and these advances are placing increasing demands on an enterprise’s IT infrastructure. The demands are not only voluminous but are also very complex with users having high expectations around fast response times, access to a large variety of services/applications and 24/7 availability of network resources.

Historically, enterprises have relied on their private infrastructures with computing platforms installed on-premises, to satisfy the needs of their users but capital and operating costs continue to increase making the growth and maintenance of computing platforms cost prohibitive. Increasing demands have also forced enterprises to look at greater economies of scale. Enterprises have turned to a flexible IT infrastructure that allows them to grow their services while keeping costs down.

Cloud computing offers a scalable, elastic and highly reliable but yet cost effective way to expand the enterprise infrastructure. It is estimated that the trend towards cloud computing will continue to grow and that the majority of enterprises will have deployed cloud computing infrastructures in the next few years. In fact according to Information Week, “2015 is the year that will mark the start of the land rush of enterprise workloads moving into the cloud.”

New big data workloads and new technologies like Hadoop are top of mind for enterprises to migrate to the cloud. Big data is now proving to be an invaluable tool as enterprises see the benefits of in depth analysis of their data to uncover customer buying patterns, sales trends and other insights that are critical to their business.

Hadoop has become the platform of choice for Big Data and has increased traction within the enterprise. An IDC study titled “Trends in Enterprise Hadoop Deployments” reports “32% of respondents indicated that their firms have existing Hadoop deployments. An additional 31% indicated that they had plans to deploy it within 12 months. And finally 36% said that their Hadoop deployment schedule could go beyond 12 months.”
Enterprise Apache Hadoop™ Deployments

Enterprises that have currently adopted Apache Hadoop™ are using one of two common deployment models: on-premises or in the cloud.

With an on-premises deployment, enterprises purchase all of the hardware and it resides in their data center. Software is installed, configured and maintained by IT administrators within the enterprise IT department. With this model, the enterprise has full access and control over the Hadoop clusters with the ability to configure the clusters whenever needed. The IT infrastructure can be extended or scaled back to accommodate changing Hadoop clusters. Data is kept private in an on-premises deployment model protecting sensitive data and ensuring that regulatory standards and compliances are met. As data volumes increase, data storage requirements also expand.

In a cloud deployment, Hadoop clusters are deployed within a service provider’s network where the configuration and administration of the Hadoop clusters are done both by the enterprise and the cloud provider. One key advantage with a cloud deployment model is that the enterprise has access to an elastic and scalable infrastructure that makes Hadoop even more cost efficient and flexible. Storage can easily scale with increased data volumes since cloud providers have the scalability to support unlimited storage. But enterprises have very little control over private data and the concern around exposing sensitive data is very real for many enterprises particularly in the financial and health care markets.

There is an additional model that is now emerging in enterprise deployments and seeing broader consideration: the hybrid cloud. The hybrid cloud provides enterprises with the advantages of both on-premises and cloud deployment models. A leading analyst firm makes a compelling statement around the adoption of hybrid cloud: “The use of cloud computing is growing and by 2016 this growth will increase to become the bulk of new IT spend....2016 will be a defining year for cloud as private cloud begins to give way to hybrid cloud, and nearly half of large enterprises will have hybrid cloud deployments by the end of 2017”

Introduction to the Hybrid Cloud

A hybrid cloud marries together on-premises deployment and public cloud deployments with all the benefits and advantages of both. For example, private data can be kept within the bounds of the on-premises infrastructure while enabling enterprises to take advantage of the elastic resources and services within the public cloud.

Hybrid clouds ease the process of migrating workloads from on-premises to the cloud. Data can be replicated across both the public cloud and on-premises to ensure business continuity. Enterprises can more efficiently deal with peak performance periods by allocating resources in the public cloud when needed without having to make additional investments in their on-premises infrastructure. Both of these advantages are critical to big data and Apache Hadoop™ deployments.

With a hybrid cloud, large datasets can be stored in the public cloud where storage systems have capacity that is readily available. Enterprises reap the benefits of technology innovations in storage that a public cloud provider offers coupled with tools for data backup and restoration.

Most importantly with a hybrid cloud, the pricing schemes are flexible with “pay as you use” models. In other words, enterprises only pay for the resources when they are needed (during quarter end or year end for example) and have access to the public cloud to extend their current infrastructure.
The table below provides a comparison of the key features and benefits of on-premises, public and hybrid cloud deployments.

<table>
<thead>
<tr>
<th>Feature</th>
<th>On-Premises</th>
<th>Public Cloud</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elasticity</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Scalability</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Protection of Sensitive Data</td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Unlimited Storage</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Table 1: Comparison of on-premises, public cloud and hybrid cloud

**Apache Hadoop™ Hybrid in Cloud Deployment Models**

As enterprises move towards the adoption of a hybrid cloud for Hadoop, there are two cloud deployment options to consider: Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS).

**IaaS**

IaaS is one of the most fundamental service models of cloud computing. IaaS such as Microsoft Azure provides computing infrastructure resources (specifically hardware) that run in a virtualized environment. In the case of big data, the Hadoop platform runs on a virtual machine and computing resources for Hadoop clusters can be pulled from numerous servers and other network resources that are located within the cloud infrastructure. The cloud provider takes care of the maintenance of the physical hardware and virtual machines ensuring that the needed resources are available to meet an enterprise’s demands.

IaaS has a lot of advantages for enterprises because it is cost effective and highly scalable. An enterprise can continue to maintain their private on-premises infrastructure and choose when they need the extra available resources with an IaaS. This allows enterprises to have business agility without incurring exorbitant hardware costs.

**PaaS**

PaaS provides platforms and software application tools that are available in the public cloud. Users can create software applications and development environments by using the features of PaaS. One of the key advantages with PaaS is that the cloud provider is responsible for the maintenance and upkeep of the software platforms or tools. In this way enterprises are assured access to the latest releases and features of their software tools or platforms of choice.

With Hadoop in a PaaS deployment such as that available through Microsoft HDInsight, enterprises have access to all the latest development tools and open source software in the public cloud and can immediately start to ingest, process and analyze data with pre-configured Hadoop clusters that require minimal to no configuration by the enterprise. With PaaS, additional analytics software applications are made available for data scientists to do in-depth analysis of their data.

Similar to IaaS, the PaaS pricing is based on a pay as you use model with subscription fees.
Enterprise Requirements for Apache Hadoop™ for a Hybrid Cloud

When planning an implementation of Hadoop in a hybrid cloud, there are some very key features that a hybrid cloud needs to deliver and enterprises must ensure that all these requirements are addressed for a successful deployment:

- **Interoperability**
  A hybrid cloud must provide full interoperability and platform deployment choices across a variety of operating systems and infrastructure platforms. A robust hybrid cloud architecture is platform agnostic and in this way, the enterprise is assured of seamless business continuity no matter which environments or platforms they are working with. Enterprise Hadoop platforms should facilitate hybrid cloud deployment models with features that enable tethering. Tethering features allow Hadoop clusters to exchange data and workloads between on-premises and cloud deployments in either a manual or automated fashion but ultimately in a fully automated and seamless manner. This drives seamless interoperability beyond operating system and software version compatibility and eases the job of the operations manager.

- **Flexibility**
  Many organizations rely on a cloud based option for their development teams and so a hybrid cloud deployment should offer the flexibility of having a development environment and a production environment within the cloud or on-premises but kept separate and distinct. This allows enterprises to manage multiple environments and also to create temporary environments for long term or short term development. For Hadoop the applications themselves need to be available in both environments.

- **Portability**
  Enterprises need to be able to move data and workloads back and forth between the on-premises cluster and public cloud cluster. As data is being moved, the key is to synchronize data sets and to ensure the consistency of information across environments. The same version of the entire Hadoop stack must be deployed in the different environments or a job execution may fail during migration. This is a critical requirement for hybrid deployments of Hadoop.

- **Scalability and Elasticity**
  The ability to scale on an as needed basis is a basic tenet of cloud computing. A hybrid cloud should allow the private IT infrastructure to expand as needed to the public cloud when workloads exceed the capacity or resources of the private infrastructure. For example, enterprises should be able to spin up Hadoop clusters on demand without the need for a large investment in additional servers.

- **Continuity**
  A loss of data or even a delay in replicating data from one environment to another can impede business continuity. One of the biggest challenges in the data center is the ability to quickly and efficiently store data without any data loss or data corruption. Enterprise grade backup and restore functions ensure data integrity and are critical in a hybrid cloud.

- **Redundancy**
  Increased globalization has forced enterprises to extend the reach of their IT infrastructure to span multiple geographic areas and is one of the key drivers around cloud adoption. In a public cloud, there is built in elasticity that allows for redundant configurations ensuring the highest levels of availability in case of disaster recovery. Incorporated in redundancy should be automated mechanisms for replication of Apache Hadoop™ clusters so that configuration settings are not lost.

- **Accessibility**
  A hybrid model for Hadoop should provide tools to migrate and replicate data. Additional applications and services are also necessary to analyze and extract valuable insights from the collected data. Ideally all these tools are available from one platform providing simplified access to IT managers, data scientists and business analysts.

- **Automation**
Backup and restoration are key components to maintaining the integrity of the hybrid cloud particularly in disaster recovery scenarios. Enterprises need the assurance of having automated backup to the public cloud scheduled on a regular basis via integrated data backup tools. Similarly, these same tools should provide data recovery that is done quickly and easily to ensure that little to no down time occurs.

**Hadoop for Hybrid Cloud Scenarios**

There are several scenarios for which a hybrid cloud model can be used for Hadoop as illustrated in Fig 1.

![Hybrid cloud scenarios for Apache Hadoop™](image)

**Application Development/Proof of Concept (POC)**

One of the key drivers in this scenario is the need for flexibility to create on demand environments for application development and testing. These environments are kept separate from the production environment to protect production data and SLA workloads from any potential development errors or load spikes.

IT managers can test out new applications, technologies or concepts on a smaller subset of resources without having to commit to on-premises infrastructure. This helps to de-risk the application development process and keep development costs predictable. Development engineers can also set up discovery sandboxes where external data is easily loaded to the cloud for manipulation and evaluation.

**Cloud Backup and Archive**

Another beneficial scenario for a hybrid deployment is one where the public cloud is used as a low cost, offsite backup for data. Data is automatically uploaded for backup depending on backup needs and then seamlessly archived in the cloud provider’s storage systems.

In a global configuration, data from data centers around the world are aggregated to one central location. Hadoop can be run local to a datacenter or can be aggregated across multiple datacenters to query the entire dataset.

**Bursting to the Cloud**
The volume of enterprise workloads tend to fluctuate depending on the time of day or time of year which makes it difficult to have exact predictions about what computing resources are required. Due to this volatility in workloads, enterprises have the need for cloud elasticity providing temporary access to needed resources. In on-premises deployments of Hadoop, clusters are configured for typical workload needs. Although there is expected volatility in data volumes, surges of user activity or data events can and do happen and enterprises must have the ability to maintain performance levels and service the organization.

In this scenario, enterprises can expand their available resources by extending the infrastructure to the public cloud and configuring clusters on-demand. Enterprises are then able to handle peak workloads in a completely interoperable environment.

With Hadoop for a Hybrid Cloud, enterprises can get additional capacity by separating out jobs. For example, an IT manager can separate analytics from reporting or recent data from archived data. Additionally, jobs can be separated according to departments or by priorities.

**Providing Hadoop for Hybrid Cloud choices**

For enterprises that want to use Hadoop as part of their Big Data solution both on premises and in the cloud, Hortonworks and Microsoft have partnered together to offer a combined solution that addresses enterprise requirements for Hadoop in the hybrid cloud. The interoperable, flexible and portable Hadoop solutions enable organizations to meet core enterprise requirements and take advantage of cloud benefits such as cost and scale as well as automated backup. Unlike other purely on-premises or purely cloud implementations, Hadoop for Hybrid Cloud gives enterprises the benefits of both, with control and flexibility of on-premises and elasticity and redundancy of the cloud. It provides a wealth of Hadoop deployment options that are fully interoperable (see Figure 2). These interoperable deployment options also open up additional advanced analytics opportunities accessible through Azure Services.

---

**The Hortonworks Data Platform**

The Hortonworks Data Platform (HDP) is the only 100% open source Apache Hadoop™ distribution available on premises and in the cloud and is the only distribution native on both Windows and Linux. HDP provides linear scalable storage and compute across a wide range of processing methods from batch and interactive to real time search and streaming. HDP provides the core components to an enterprise data platform: data access, data governance, data management, security and operations. Data governance with Apache Falcon, delivered as part of HDP, is critical for a hybrid cloud deployment and facilitates data
movement from on-premises to the cloud. HDP integrates with existing systems in the enterprise data center with minimal changes to the overall infrastructure. HDP has been deployed in on-premises, private and public and infrastructures.

Hadoop is part of a Modern Data Architecture that can span both in an on-premises or cloud deployment. The Modern Data Architecture enables the adoption of a hybrid architecture because enterprises can pick and choose which components belong on-premises or in the cloud.

Figure 3: The Modern Data Architecture

Azure and Azure HDInsight

Microsoft offers a cloud platform with Microsoft Azure that provides a collection of integrated services including compute, storage, data, networking, backup and restore. Microsoft Azure has been ranked by a leading analyst as an industry leader for both IaaS and PaaS. This combination gives enterprises the best of both worlds and enables the build out of hybrid cloud deployments without added complexity. Microsoft Azure offers flexible pay as you use models that allow enterprises to expand their capacity as needed.

Microsoft HDInsight deploys and provisions Hadoop clusters in the cloud, providing a software framework designed to manage, analyze and report on big data. Azure HDInsight is built using the Hortonworks Data Platform.
Hadoop for Hybrid Cloud: Benefits and Advantages

The Hadoop for Hybrid Cloud solution has been architected to give enterprises a wide breadth and depth of features to make the migration to a hybrid cloud easy and cost effective (see Figure 4).

Some of these features include:

- **Interoperability, Flexibility and Portability**
  Through the joint engineering partnership, Microsoft and Hortonworks has built HDP Windows native distribution and contributed back to the Open Source community. Enterprises have a choice of operating systems and infrastructure deployment platforms with seamless interoperability.

- **Mission critical capability and business continuity at cloud scale**
  Microsoft and Hortonworks enable on-premises Hadoop customers to gain the benefit of mission critical scale and redundancy with Microsoft Azure cloud. Customers can backup their on-premises data using Falcon as well as leverage the elasticity of the cloud to scale-out during peak demand.

- **Accessibility to Cloud analytic services**
  By leveraging the power of Microsoft Azure, customers have access to advanced analytics and data services available to them from the cloud. New types of data can be processed with Microsoft HDInsight. Enterprises can leverage the power of Microsoft Azure Machine Learning to predict future trends or behavior without needing a deep data science background. Or Microsoft Azure Data Factory can be used to orchestrate and curate data through simple, fault tolerant data pipelines. All of this and more are available in Microsoft Azure without hardware to install or software to maintain.
Conclusion

The use of Hadoop in hybrid cloud deployments is on the rise and enterprises need a robust cloud platform to lead them into the next generation of cloud infrastructures.

Through their deep engineering partnership, Hortonworks and Microsoft have made Hadoop for Hybrid Cloud available so that migration to a hybrid cloud architecture can be done efficiently and seamlessly. The integrated solution gives enterprises the flexibility, scalability, backup and analytics capabilities that they need all on one platform.

For more information on Apache Hadoop™ in hybrid clouds go to:

http://hortonworks.com/partners/microsoft-hybrid/

For more information on Microsoft Azure and HDInsight go to:

http://azure.microsoft.com/en-us/services/hdinsight/
http://azure.microsoft.com/en-us/

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. The Hortonworks Data Platform provides an open platform that deeply integrates with existing IT investments and upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit http://www.hortonworks.com