GIGAOM RESEARCH

Managing Hadoop: what's the best approach?

George Anadiotis

November 25, 2014

This report is underwritten by Zettaset.

TABLE OF CONTENTS

Executive summary

Currently, cloud-based Hadoop clusters are receiving a lot of publicity while most of the Hadoop action is on-premises (on-prem), but an emerging option promises to combine the best of both approaches by adding a management layer to Hadoop on-prem.

This report will investigate each of the three options on this spectrum, analyze the benefits and disadvantages of each, and inform a sound strategy for organizations looking at Hadoop adoption or changing their current process.

Key findings:

- Organizations leveraging on-prem Hadoop can control their deployments and apply optimal configuration for their workloads. They eliminate recurring infrastructure costs and any worries about data leaving their premises. On the other hand, building and maintaining this kind of in-house infrastructure requires an up-front investment as well as expertise, and using it optimally is not always possible.

- Organizations leveraging Hadoop in the cloud can achieve insight more quickly because they benefit from elasticity and do not have the burden of building and maintaining the infrastructure and developing the expertise required for supporting in-house Hadoop clusters. On the other hand, costs can accumulate quickly, networking and latency usually prevent cloud being optimal for large workloads, and the organization gives control to third parties.

- Organizations can benefit from a management layer on top of Hadoop that enables them to have the best of both approaches, but this layer must include advanced enterprise security and systems-management features.

DIY Hadoop on-prem: all about control

Hadoop was built for scaling cost-effective commodity infrastructure. However, as the rate of data growth does not show any signs of slowing, not every organization can afford to identify and hire the specialized IT talent required to deploy a Hadoop environment successfully. Other well-funded and highly resourced organizations were early to assign dedicated resources to their Hadoop clusters, taking a do-it-yourself (DIY) approach to building and expanding them at a constant pace.

Many organizations have expanded their cluster-management skills with their Hadoop deployments, either leveraging existing skills or building them as they go. They have learned to exercise the ultimate in control in their clusters, but the truth is that this control comes at a cost.

Data volumes and their associated bandwidth limitations motivate on-prem deployment. Hadoop's MapReduce architecture requires storing massive amounts of data close to compute nodes as well as moving data among nodes to perform computation in the reduce step. The bandwidth requirements are considerable and handled better by a local network that does not have the topology and connectivity limitations that cloud nodes do. Furthermore, data in local nodes – as opposed to data in typical cloud nodes – is not transient, which poses a great burden on cloud deployments, as data must be moved to nodes every time it is instantiated. Otherwise a cloud storage service would have to be leveraged, incurring additional cost and complexity.

Cost is a significant factor. While building an on-prem cluster may require significant upfront investment, over time, the total return on investment (ROI) is favorable compared to the cloud option. Buying nodes instead of renting them makes more sense long term, if the investment in a Hadoop cluster is a strategic one.

Another factor to consider is security, which still inhibits cloud adoption for many organizations. Cloud vendors are addressing these concerns, but in some cases, regulations prohibit sensitive data leaving the premises, so on-prem processing is the only option.

Building a Hadoop cluster is no easy feat. It represents a significant investment that organizations must make, including costs for infrastructure procurement, personnel training, and operational costs. Data-center tasks incur up front and recur with hardware and personnel costs. Sufficient cluster nodes and networking resources must be in place for an on-prem cluster to function properly.

Estimating what is "sufficient" can be difficult, as cluster workloads typically vary. So, organizations can have either a cluster that can handle small workloads and chokes when volume increases or a cluster that can handle big workloads and remains underutilized most of the time. A private cloud can mitigate this typical lack of elasticity, but it comes with additional cost.

Appropriate personnel must be hired and trained to operate and maintain an on-prem cluster. Keeping up-to-date with an evolving ecosystem such as Hadoop requires significant investment and may develop into a headache for organizations not strategically interested in building a dedicated Hadoop infrastructure team. These companies must rely on professional services to deploy and manage clusters using branded Hadoop distributions from vendors such as Cloudera, Hortonworks, and MapR.

A handful of large organizations in the internet domain were pioneers of in-house DIY Hadoop deployment. They have had to unify data coming from many different sources, with overwhelming volumes that continue growing at an ever-increasing pace. These organizations have grown alongside Hadoop and use it to analyze clickstreams combined with other data to gain insights into current and future customer behavior. They have dedicated machines and personnel to run their clusters and many of them have also built their own custom layers of functionality on top of Hadoop. With few budget or resource constraints, these pioneering organizations are in total control of their Hadoop installations and can operate them optimally. However, the vast majority of enterprises cannot afford this level of control.

Hadoop in the cloud: lifting the burden

Although a few select and forward-looking organizations have grown their Hadoop expertise as the platform has developed, they represent a small percentage of all organizations that would benefit from Hadoop. The rest find building or recruiting their Hadoop skills base from scratch daunting, as the goal usually is to get immediate actionable results. Even organizations with in-house expertise find the option of turning to a cloud-based third party to install and maintain their Hadoop clusters compelling. Off-loading the burden of maintenance while enjoying the benefits of cloud elasticity is tempting.

The first reason is ease of use. Cloud-based Hadoop managed services mean no more infrastructure procurement, tedious installation, or configuration. Fully managed cloud environments promise that everything is pre-configured and ready to use.

Cloud-based Hadoop clusters are self-service. Someone else has hired personnel and trained them to cope with set-up details, so organizations can sit back and enjoy the benefits. Elasticity is also important. Cloud-based Hadoop clusters can scale on demand, so organizations can get exactly what they need, when they need it, and give it back when they have finished using it.

But even in cases when Hadoop cluster management is completely outsourced to a cloud provider, the company must have a minimum familiarity with Hadoop to submit jobs. The most obvious argument against using a managed cloud-based Hadoop cluster is cost. The upfront cost is minimal, but recurring costs can quickly accumulate beyond the upfront investment. (Consider, however, that while cloud costs are recurring, for planning purposes they are typically more predictable than costs associated with on-prem deployments, especially DIY approaches that require specialized personnel.)

Other cloud issues include network latency and transient storage. Not much data exists at present on the average Hadoop workload, but a recently published survey calculated that input, shuffle, and output sizes of all jobs in the workload ranged from 80 Terabytes to 18 Petabytes – a lot of data to be moving around in the cloud. So organizations can either utilize permanent storage nodes – adding an extra layer to Hadoop Distributed File System (HDFS) and paying the associated price – or move around their data each and every time new nodes are spun.

But even acquiring permanent storage does not mitigate network connectivity in the cloud to make it appropriate for running a Hadoop cluster. Communication between nodes will most likely not be as fast as needed for optimal performance unless very careful consideration is given to fine-tuning a cloud-based cluster.

Also, organizations must understand that by signing up to a cloud-based managed Hadoop cluster solution, they are embracing their partner of choice and entrusting that partner with important data and skills. In many ways, this is similar to outsourcing database management, which is a strategic decision every organization must comprehend fully before signing up.

Self-service is not easily realized today because of the high reliance on necessary professional services to deploy Hadoop from branded distribution vendors. Likewise, the traditional cloud benefits of infrastructure elasticity and scalability cannot be realized because overall, existing Hadoop distributions provide very low levels of software automation and require a lot of resources to support time-consuming, error-prone, and costly manual processes.

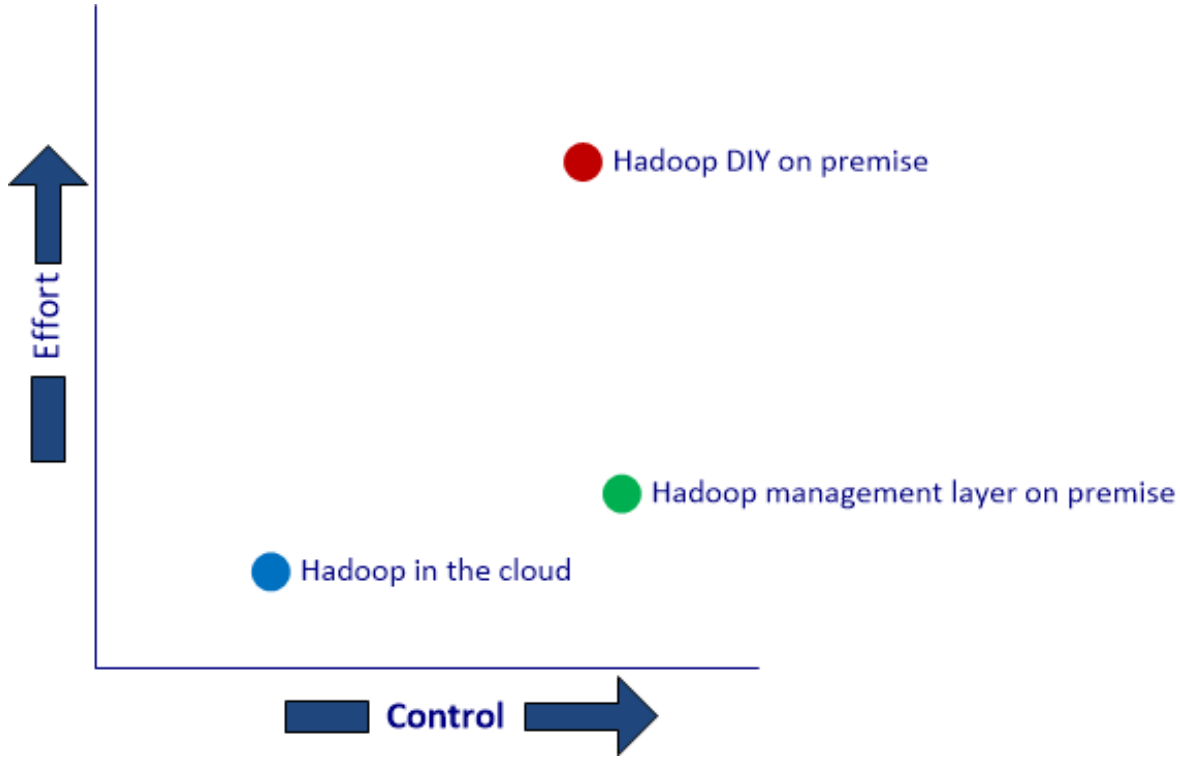
One media company that has invested in a cloud-based Hadoop cluster deployed on Amazon's EMR has adopted Amazon's S3 storage layer instead of Hadoop's own HDFS storage layer so that it can overcome the transient storage issue. It has also built a management layer on top of its cluster, thus streamlining and controlling access to it. Although impressive, the average organization can't easily implement this strategy. And despite the attention and resources that have gone into it, the strategy still gives away a level of control.

The best of both worlds: the importance of being manageable

Today, even well versed organizations using Hadoop are struggling to make it ready for the enterprise data center, not because they are lacking in expertise, but because many features associated with more mature database technologies, such as Relational DBs, are not present in the current incarnation of the Hadoop ecosystem. This functionality requires an additional layer on top of Hadoop.

Hadoop was designed with commodity infrastructure in mind and is able to cope with the computational challenges it imposes. The secret to making Hadoop manageable doesn't lie in the infrastructure used to deploy it, but rather in the added-value features that must be included in its current ecosystem. Hadoop was never built to address scenarios that go beyond a closed group of experts with access to data. Today, Hadoop is touted as the source of truth for data applications that span the enterprise (the data lake metaphor), but it was never designed to be that.

Level of control vs. required effort for different Hadoop management options



Source: Gigaom Research

Security is the single most important feature that must be addressed for Hadoop to become enterprise ready. It spans three key areas: fine-grained data access, encryption, and audit trails.

Fine-grained access control. Hadoop accesses data by placing files in its HDFS storage layer. HDFS is a file system with permission-based access similar to one used by UNIX file system. Files are stored in folders and users belong to groups. Files and folders have three sets of permissions: read, write, and execute. And they are organized into three levels: owner, group, and world.

While this management scheme is simple and familiar, it is not adequate for enterprise use. HDFS files are the equivalent of database tables that a multitude of users can access. The problem is that the permission scheme can only be applied to entire files, not to the row level and column level. The scheme needs a more finely grained level of permissions.

Encryption. An API in out-of-the box Hadoop can enforce encryption on the application level, but no file-level encryption option is available for HDFS. Although encryption can be added on the application level, implementing it is a very effort-intensive task. Some vendors offer standalone encryption implementations that can be plugged into Hadoop. However most of these solutions are partial, as they only address data at rest and not data in motion. In addition, typically these solutions do not address other areas of Hadoop security; they are implementations of a specific encryption algorithm.

Audit trails. Although HDFS can log all file system access requests, the implementation of this feature is rather low-level (using log4j logging). In addition, auditing must be applied not only for HFDS, but also for services and data. Audit trails must exist to locate who has accessed data and services (and when) and auditing is needed to track data lineage – where datasets originate and when, where, and how they have been processed. This audit helps track results and adds to their credibility.

The other set of features that must be addressed by a layer on top of Hadoop is manageability: simplified, automated deployment and configuration, high availability/automated service failover, and monitoring.

- **Automated deployment and configuration.** Hadoop cluster deployment is not an easy task. Hadoop and its components must be installed with all their dependencies across a number of nodes and communication must be established, tested, and optimized. Usually administrators with a high degree of specialization carry out this task, whether it takes place internally or whether external resources are leveraged for it.

In either case, this is an expensive, labor intensive, and error-prone process to perform and use. Ideally, a software layer should automate such tasks. It should be able to take the Hadoop installation files and the nodes available for the cluster as input, and output a ready-to-operate Hadoop cluster.

- **High availability and monitoring.** The task of configuring a Hadoop cluster does not end with successful deployment. As workloads start coming in, node utilization and health must be constantly monitored so that potential downtime can be avoided by directing workloads to less heavily utilized nodes, and nodes that experience downtime can be restored online as quickly as possible to minimize disruption. This is also a labor-intensive task that is typically carried out by specialized personnel, whether internally or beyond organizational boundaries. Even though this is a hard task to automate fully, it would be extremely beneficial to do so as the associated resource cost would diminish.

  The strategy of utilizing a layer above an internal Hadoop cluster can be viewed as combining the best of both worlds – retaining the level of control that on-prem solutions give, while, similar to managed-cloud solutions, removing much of the burden that comes with them.

  An organization that applies this is a biosciences company that needs to manage biomedical data coming from multiple sources and to combine genomics data with payer-provider data. Utilizing Hadoop with an added management layer, this company can leverage its capabilities for advanced clinical trials and push results out to physicians quickly. Managed Hadoop infrastructure can also perform next-generation genome sequencing for large numbers of people at low cost.

  But, the choice between Hadoop on-prem and Hadoop in the cloud is not as clear as black and white. Other factors influencing Hadoop's on-prem gravity include data volumes, security concerns, bandwidth limitations, and overall ROI. Other factors, like self-service, ease-of-use, and the lack of networking and storage-management burden make the cloud appealing nonetheless.

Key takeaways

Organizations with different needs, strategies, and degrees of sophistication are leveraging Hadoop. Each strategy has advantages and repercussions. Some pioneers that have strategically invested in Hadoop are taking the DIY approach, as they have the resources and vested interest to do so. They have maximum control, but at a high price. Other organizations more interested in getting immediate results and not willing to invest in infrastructure and personnel are opting for cloud-based Hadoop offerings. This option can minimize time to insight, but gives away control. And, some organizations adopt a managed on-prem approach that promises to take away the pain associated with building and maintaining a DIY Hadoop cluster on-prem, while letting the organization retain control.

- DIY Hadoop cluster management offers a great deal of control, but it comes at a cost that few organizations can afford. The benefits of optimal cluster configuration, better overall ROI, and retaining control of data can be offset by upfront investment, training/hiring personnel, and dependence on professional services.

- Hadoop in the cloud trades control for simplicity, but over time, costs accumulate and strategic assets may be at risk unless appropriate levels of data security are in place. The benefits of pre-configured clusters, self-service access, and elasticity can be offset by limitations in networking, latency, transient storage, and recurring costs.

- On-prem managed Hadoop promises to balance the best of both approaches, providing more of the required enterprise features for running Hadoop than are currently available in cloud environments, while typically being more cost-effective than DIY approaches.

- Two qualities all organizations should look for in their managed Hadoop on-prem solution are security (e.g., fine-grained data access, encryption, and audit trails) and manageability (e.g., automated deployment and configuration, high availability, and monitoring).

About George Anadiotis

George Anadiotis has been active in IT since 1992, having worn many hats and juggled many balls. As a ninja programmer, a lead architect, a team manager, a trusted consultant, an entrepreneur, and an analyst, he has provided services to the likes of KLM and Vodafone, built and managed projects and teams of all sizes and shapes, and become involved in award-winning research.

The common threads that span these activities are integration, modeling, and analysis, be it on the application, data, process, or business level. He enjoys researching, developing, applying, and evangelizing cutting-edge concepts and technology, was among the pioneers in enterprise-application integration, and is among the pioneers in big-scale data integration.

About Gigaom Research

Gigaom Research gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports, and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, Gigaom Research addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: research.gigaom.com.