Deploy an OpenStack private cloud to a Hadoop MapReduce environment

Steven C. Markey (steve@ncontrol-llc.com)
Principal
nControl

15 October 2012

Cloud computing and big data are changing the enterprise. This article explains why it's necessary to assimilate these new technologies to achieve a maximum return on investment on your analytics platform. Read on to gain an understanding of what a private cloud is, what cloud computing and big data technologies consist of, what synergies they present, and how to deploy them.

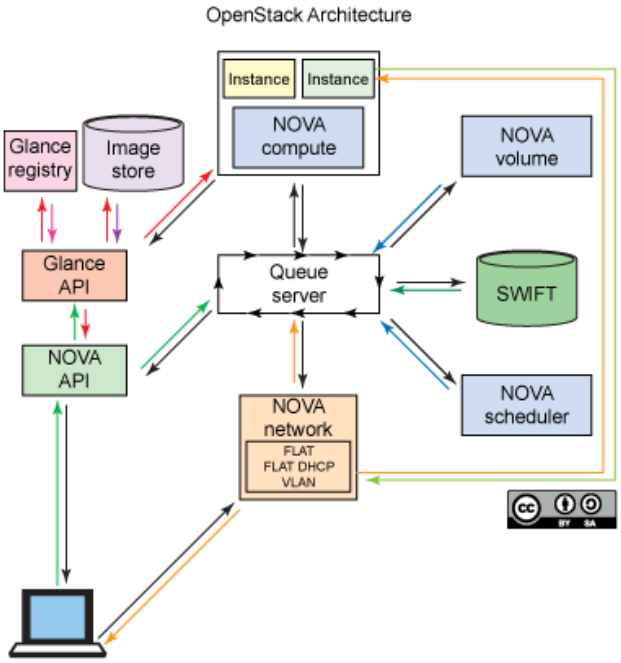
As organizations leverage both cloud computing and big data technologies in the enterprise, it's wise to consider combining these tools. In doing so, an enterprise reaches an optimal level of analytical processing capability while leveraging the rapid elasticity and single tenancy of a private cloud. This article helps you gain an understanding of what cloud computing and big data technologies consist of, what a private cloud is, what Apache Hadoop is, the synergies they present, how to deploy these technologies, and the challenges they present.

Private cloud computing primer

A *private cloud* is an internal cloud computing deployment within the enterprise in which the organization leverages the nuances of cloud computing technologies inside the data center. These nuances include rapid elasticity, resource pooling, on-demand provisioning, and automated management. To incorporate all of these attributes internally, most organizations use an open source cloud distribution, such as OpenStack or CloudStack.

OpenStack is the most prevalent of the open source cloud distributions and includes controller, compute (Nova), storage (Swift), messaging queue (RabbitMQ), and networking (Quantum) components. Figure 1 provides a diagram of these components (without the Quantum networking component).

Figure 1. Components of an OpenStack distribution



These components are bound together to deliver an environment that allows for the dynamic provisioning of compute and storage resources. From a hardware standpoint, these services are spread out over many virtual and physical servers. As an example, most organizations deploy one physical server to act as a controller node and another to serve as a compute node. Many organizations choose to parse out their storage environment onto a dedicated physical server, as well, which in the case of an OpenStack deployment would mean a separate server for the Swift storage environment.

Big data primer

Oracle defines *big data* as an aggregation of data from three sources: traditional (structured data), sensory (log data, metadata), and social (social media). Big data is often stored using new technology paradigms, such as non-relational, distributed databases like NoSQL. There are four types of non-relational database management systems (NRDBMSs): Column based, key-value, graph, and document based. These NRDBMSs aggregate the source data while analytical programs, such as MapReduce, analyze the aggregated information.

A traditional big data environment includes an analytical program, a data store, a scalable file system, a workflow manager, a distributed sorting and hashing solution, and a data flow programming framework. The data flow programming framework often used for commercial applications is Structured Query Language (SQL), while for open source distributions, it is an alternative to SQL, such as Apache Pig for Hadoop. On the commercial side, Cloudera provides one of the more stable and comprehensive solutions, while Apache Hadoop is the most prevalent of the open source Hadoop distributions.

The use of the Apache Hadoop distribution is common because of the variety of components you can use, including the Hadoop Distributed File System (HDFS—a scalable file system), HBase

(database/data store), Pig, Hadoop (analytics), and MapReduce (distributed sorting and hashing). As Figure 2 shows, Hadoop tasks are broken down into nodes, while MapReduce tasks are broken down into trackers.

Figure 2. HDFS/MapReduce layer composition

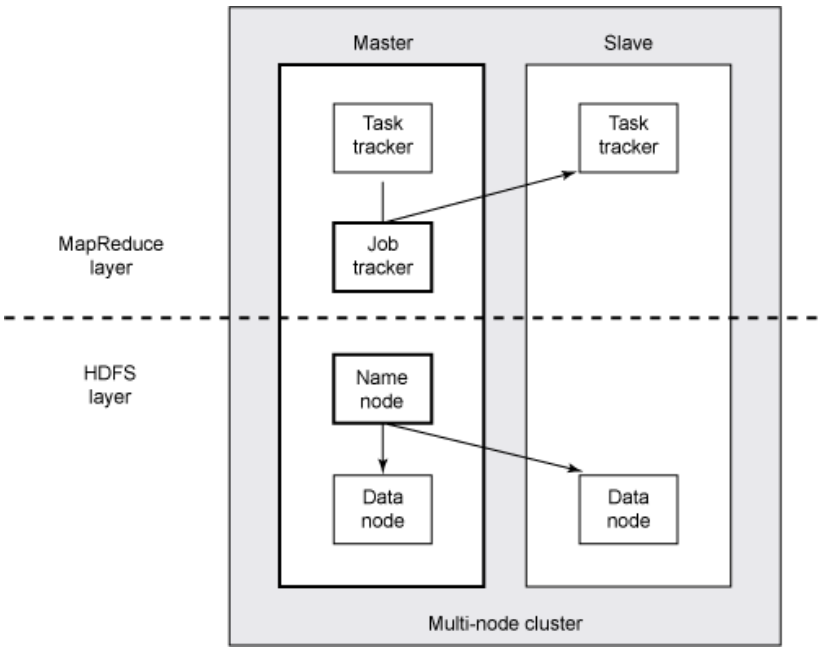


Figure 3 shows how MapReduce performs its job, which is by taking input and executing a set of split, sort, and merge actions before presenting the sorted and hashed output.

Figure 3. High-level MapReduce diagram

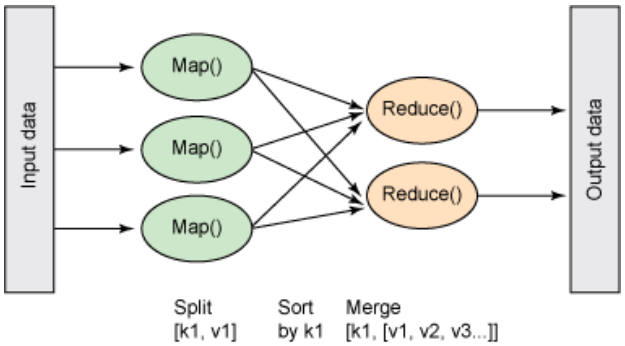
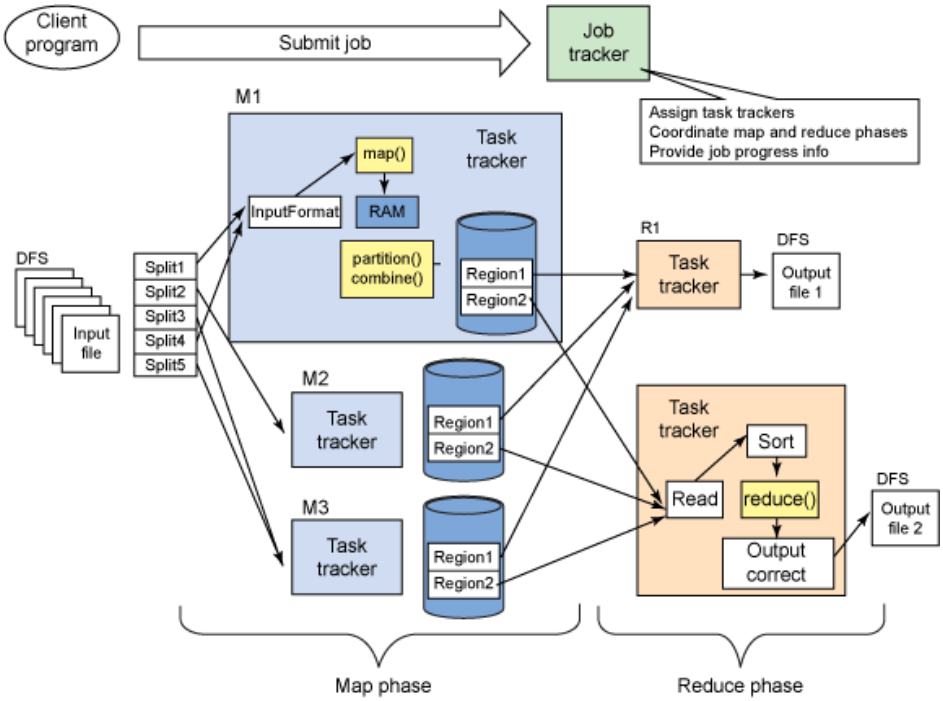


Figure 4 illustrates a more complicated MapReduce job and its components.

Figure 4. MapReduce data flow diagram



Although more complicated than traditional analytical environments (such as IBM® Cognos® and Satori proCube online analytical processing), Hadoop MapReduce deployments are scalable and cost-effective.

Bringing it all together

Big data and private cloud environments are useful in their own right; however, when they are joined together, the benefits an organization experiences are substantial. Although the environment will be complex, an enterprise will see substantial synergies by joining an OpenStack private cloud with an Apache Hadoop environment. The next sections walk through how an organization can integrate private cloud and big-data technologies.

Swift, Apache Hadoop, and MapReduce

A common deployment model for big data in a private cloud environment is to deploy OpenStack's Swift storage technology joined to an Apache Hadoop MapReduce cluster for processing. The advantage of using this architecture is that an organization will have a scalable storage node to handle its ever-amassing data. Per IDC, the annual data growth rate is 60%; so, this solution will handle the challenge of ever-growing data while allowing an organization to concurrently launch a pilot for deploying a private cloud.

This deployment model is best used in an organization looking to pilot a private cloud through a storage pool while concurrently using in-house big data technologies. Best practices dictate that you deploy your big data technologies to your production data warehouse environment first, then build and configure your private cloud storage solution. When you have your Apache Hadoop MapReduce technologies successfully joined to your data warehouse environment and your

private cloud storage pool is built and working correctly, then you can integrate the private cloud storage data with the scheduled Hadoop MapReduce environment.

Swift, Cloudera's Distribution for Apache Hadoop

For those organizations that are reticent to engage in using big data from the ground up, there are big-data appliances from solution providers like Cloudera. Cloudera's Distribution including Apache Hadoop (CDH) solution allows organizations to forego having to invest in procuring or training staff in every nuance of Hadoop and thus may allow for a higher return on investment (ROI) from a big data standpoint. This route is particularly appealing to those organizations that do not have the skill set in either big data or private cloud and would like to slowly and iteratively incorporate this technology into their portfolio.

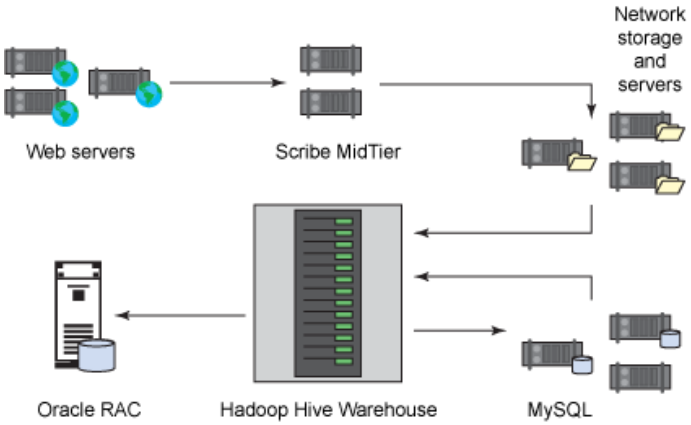
Big data and cloud computing are relatively new technologies that many organizations would like to embrace for cost savings; however, they are hesitant to embrace these new technologies in full. By leveraging a big data software distribution that their vendor supports, these organizations will have peace of mind while concurrently learning how to use this technology to their advantage. In addition, firms may see a higher utilization with this investment if it's used to analyze large data sets that can be best managed by a private cloud storage node. To best incorporate this strategy into the enterprise, first install, configure, and administer CDH to analyze the organization's data warehousing environment, and then add the data stored in Swift to the fray.

Swift, Nova, and Apache Hadoop MapReduce

For those organizations looking for an advanced level of flexibility, scalability, and autonomy within their big data environment, they can leverage the native abilities of the open source offerings provided by Apache and OpenStack. To do so requires that the enterprise leverage both technology stacks to the greatest extent possible, which in turn requires a different mentality to architecting this environment than what has been mentioned in the solution examples above.

To have a fully scalable and flexible big data environment, it must run on a private cloud environment that provides both storage and compute nodes. To do that, the organization must build the private cloud first, then add big data to the equation. So, at this point, Swift, Nova, and RabbitMQ are certainly needed as well as controller nodes for managing and maintaining the environment. However, the question is whether the organization needs to segment sections of this environment for different systems (for example, non-big data virtual machine or guest instances), business units, and departments. If the firm is going to leverage a private cloud across the board, Quantum should be added to the equation for segmenting the different environments from a networking standpoint (Figure 5).

Figure 5. OpenStack architecture



When the private cloud environment is set up and tested, incorporate the Apache Hadoop components into it. From this end, Nova instances can be used to house NoSQL or SQL data stores (yes, they can coexist) as well as Pig and MapReduce instances; Hadoop can be on a separate, non-Nova machine for processing. It is hoped that in the near future, Hadoop will be able to run on a Nova instance, making the private cloud self-contained on all Nova instances.

GFS, Nova, Pig, and MapReduce

From an architectural perspective, there are other options than using OpenStack's Swift for scalable storage. This example uses Google File System (GFS) with Nova and Apache Hadoop components—specifically, Pig and MapReduce. This example allows an organization to focus on developing a private cloud compute node only for computational processing while leveraging Google's public storage cloud as a data store. With this hybrid cloud, the organization can focus on a core competency of computational processing while a third party focuses on storage. This model can leverage other providers' storage solutions, such as Amazon Simple Storage Service; but before using any external storage, organizations should build this solution with extensible file system (XFS) internally and test it properly before extending it to a public cloud. Furthermore, depending on the sensitivity of the data, the enterprise may want to use data-protection mechanisms, such as obfuscation, de-anonymization, encryption, or hashing.

Tips and tricks

When incorporating cloud computing and big data into the enterprise, it's important to build your staff's skill set within each technology platform. When your staff understands these technologies, build a lab to test joining these platforms together. Because there are many different components, be sure to go down the proven routes highlighted above for implementation. Furthermore, should the firm hit any snags when trying to join these paradigms, move on to another alternative after a set number of attempts. Examples of alternatives include appliances and hybrid cloud.

Roadblocks and land mines

As these technologies are fairly new, most organizations will want to test them with existing resources before realizing a significant capital expenditure (CapEx). However, without properly

budgeting and staffing for assimilation of these technologies into the enterprise, pilot and testing could end in disappointment. Also, absent a complete private cloud roll-out, the organization should implement big data first into the enterprise, then the private cloud.

Finally, it is imperative that the organization devise a strategic roadmap for its private cloud and big data plans. A successful deployment will lead to many demands for additional analytical "jobs" that may hamper processing. To remediate this risk, an iterative project management methodology should be used to roll out these technologies to the enterprise via phased deployments to the business units and departments.

Conclusion

Cloud computing and big data are upon us, and it behooves your organization to determine how these technologies may benefit the company, such as cost savings from a CapEx standpoint or enhanced processing. Your enterprise should test these systems independently, and then look to assimilate them into the enterprise by deploying them iteratively. By doing so, the company can see a solid ROI that prepares the organization for the future.

Resources

Learn

- See OpenStack's Starter Guide for more information on the OpenStack architecture.
- See Michael Noll's blog post for more information on the delineation of HDFS/MapReduce.
- See Ricky Ho's Hadoop MapReduce Architecture for more information on Hadoop MapReduce.
- See Christopher Olston's Yahoo! Hadoop Explanation for more information on using OpenStack and Apache's Hadoop MapReduce.
- See Borthakur's and Shoa's Hadoop and Hive Development at Facebook for more information on Apache Hadoop deployment.
- See Oracle's Oracle: Big Data for the Enterprise for more information on Hadoop-based appliances.
- Explore developerWorks Cloud Computing, where you will find valuable community discussions and learn about new technical resources related to the cloud.
- Stay current with developerWorks technical events focused on a variety of IBM products and IT industry topics.
- Follow developerWorks on Twitter.

Get products and technologies

- Evaluate IBM products in the way that suits you best: Download a product trial, try a product online, use a product in a cloud environment, or spend a few hours in the SOA Sandbox learning how to implement service-oriented architecture efficiently.

Discuss

- Get involved in the developerWorks community. Connect with other developerWorks users while exploring the developer-driven blogs, forums, groups, and wikis.

About the author

Steven C. Markey

Steve Markey is a consultant, adjunct professor, and the current president of the Delaware Valley (Greater Philadelphia) chapter of the Cloud Security Alliance (CSA). He holds multiple certifications and degrees and has more than 11 years of experience in the technology sector. Steve frequently presents on information security, information privacy, cloud computing, project management, e-discovery, and information governance.