Review of Hadoop Platforms & Distributions

Abstract

This paper provides a detailed assessment of different cloud computing platforms and Hadoop distributions, and offersa guide for organizations to choose the most ideal platform. We will review some of the top Hadoop options that are available in the market place and discuss the best fit scenario of these options.

Introduction

We are moving at an incredible pace to a world of Internet of Things (IoT). This means that, a large volume of structured and un-structured data is being generated by a myriad of location aware and wi-fi enabled devices. Data is also being generated through the billions of transactions and activities on web sites and social media sites. All this data, referred to as “Big Data”, is a treasure trove of information and knowledge. We have moved from the problems of how to capture, store, secure, and protect data, to the problem of how to analyze the volumes of existing data for information that can be used to improve business processes. When the right technologies can be deployed to mine and distill this data, it can provide insightful information for decision makers at all levels of an organization.

Relational Database Management Systems (RDBMS), like Oracle and Microsoft SQL Server have been at the center of most systems for storing data generated by transactional applications and for data aggregations like data warehouses. While RDBMS excel for those applications, recent history has shown that some of the constraints on data consistency and availability must be relaxed to allow for management and analysis of data in the petabyte range and at Internet scales. This is because of several factors – the volume of data that is now available is more than RDBMS were designed to handle; traditional APIs and ETL mechanisms cannot handle the velocity with which the data can arrive in the new world; the new data stream consists of structured and unstructured formats that are not easily handled by the Table structure (rows and columns) of a Relational Data Base. A new architecture, called Hadoop, which can meet the new requirements, is fast becoming the choice architecture for dealing with Big Data. Hadoop at its core is redundant file system (HDFS) and a frame work for parallel processing (map reduce) that stores and processes information in a distributed fashion. The computational capacity of a Hadoop cluster is in principle only limited by the number of nodes that are part of the parallel cluster. Hadoop allows for analyzing information without worrying about scalability.

However Hadoop is not for the faint hearted. Hadoop is open source software that was originally created by the Apache Foundation. To set up a Hadoop environment and to get value from the effort, an organization has to invest in hardware, personnel and training. Personnel with knowledge of server and network hardware, operating system administration, and application development are needed. Luckily help is on the way for the infrastructure, software and tools. With the gaining of the popularity of the Hadoop software, several companies such as Cloudera, Horton Works, Intel have come out with Hadoop distributions under various brand names. These distributions come bundled with the core software, set up tools, monitoring tools and software support to make the process of installation, configuration and ongoing support easy. Several hardware and software vendors such as IBM, Dell, HP, and Microsoft are also jumping into the fray with their own appliances, which are a bundling of the hardware and software, which can be readily deployed to solve a particular use case of the Big Data. Several cloud providers such as Amazon and Google are providing their cloud platform as a service to standup Hadoop environments. These developments now offer organizations phenomenal options and the opportunity to focus on the application of the Hadoop technology to handle their Big Data as opposed to being worried about setting up the environment.

In this article we will review some of the top Hadoop options that are available in the market place and discuss the best fit scenario of these options

Cloud Computing Platforms and Hadoop

The recent demand for Hadoop combined with the lack of expertise in managing large, parallel systems has resulted in a many Cloud based solutions, including Infrastructure-as-a-Service (Iaas), Platform-as-a-Service (PaaS) and Hadoop-as-aService (HaaS) providers. These services enable companies to meet business analysis objectives while minimizing infrastructure and Hadoop management issues. Although some activities are minimized, it is important to note that the system administration, monitoring, application development and debugging, etc. are all still the responsibility of the customer.

This market is still maturing and rapidly evolving. Providers offer a wide range of features that can be challenging and time consuming to carefully evaluate before committing. Vertex Computer systems can help navigate this landscape and provide services for developing a Road Map for Big Data BA/BI In the Cloud.

Before deploying workloads in a public cloud and moving business critical data into the Hadoop cluster, the security criteria, Hadoop distributions, and supported analytic tools (e.g. Tableau, R, SPSS, SpotFire, Jaspersoft, etc.) need to be evaluated. Highly secure and strictly compliant enterprise applications require special attention and architecture. Where the data is generated should also be a deciding factor. Cloud storage is slower and more expensive for data sets that continue to grow. Hadoop clusters in the cloud make a lot of sense when the data itself is generated in the cloud (e.g., analysis of Twitter). For real-time, customer-facing systems with data coming from multiple sources, customers will likely need to build Hadoop out in a physical facility with appropriately sized network bandwidth and latency to support the application.

Implementing and managing a Hadoop cluster efficiently and cost effectively is not practical for small and medium sized businesses. This makes cloud based platforms an attractive option. Conversely, for large scale, data driven companies like the Internet giants, it is more economical to do in-house. Other large enterprises would benefit from using the cloud for a onetime workload, such as in the often cited 2007 case where the New York Times used the power of hadoop and Amazon EC2 for just one day to convert a large number of TIFF documents to PDFs in their digitization initiative. For businesses of any size the cloud based platforms are a great option for application development environments.

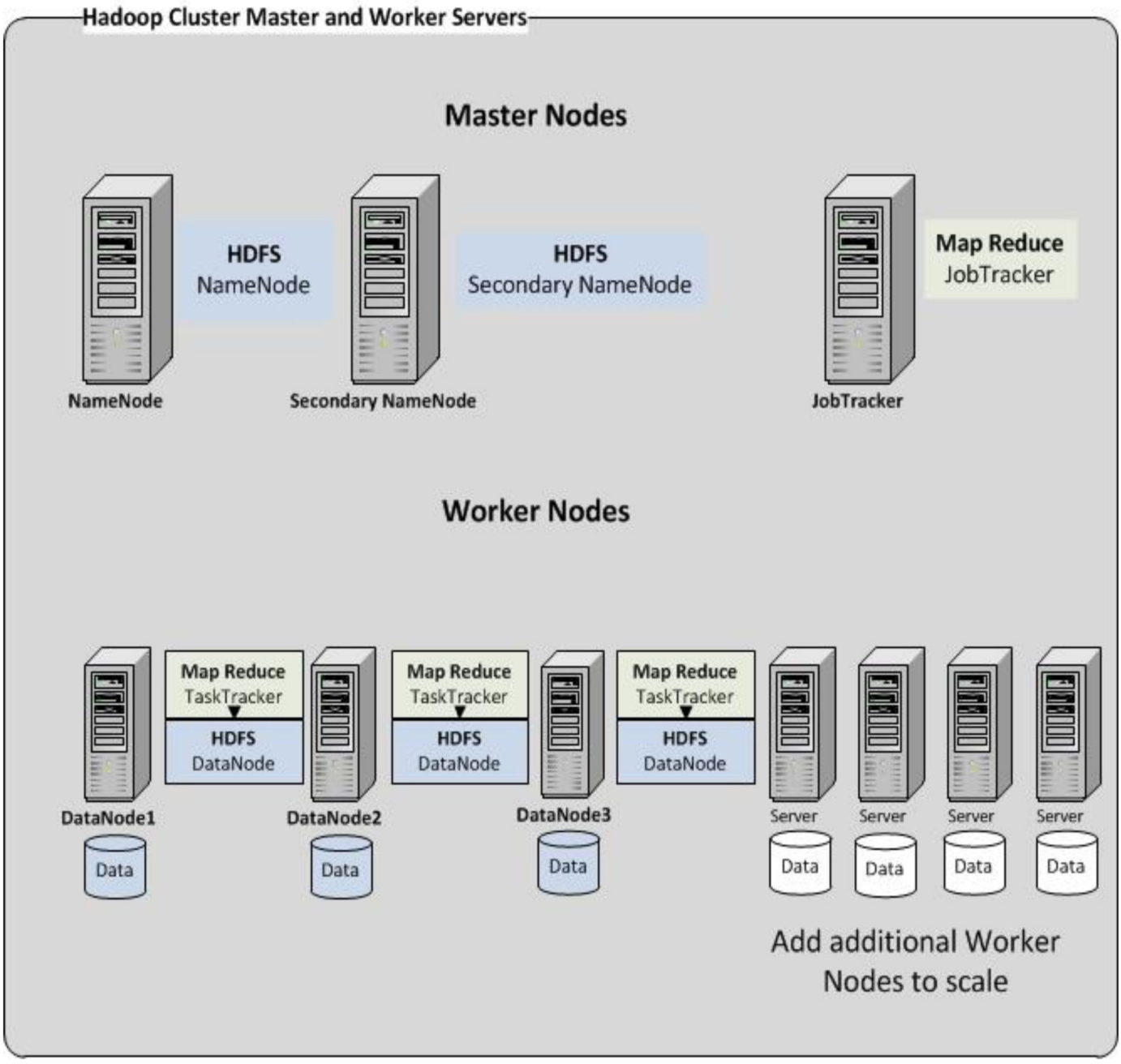
The myriad of cloud platform offerings run from the well-established front runners including Amazon (Elastic Map Reduce for HaaS), Google (BigQuery for PaaS), and RackSpace, to more recent entries like Microsoft Azure, IBM Cloud Softlayer, HP Helion Public Cloud, down to the niche players, such as Altiscale (Haas), Xplenty (Haas), Qubole(HaaS), Joyent, and CSC(Infochimps for Haas). It is important to note that some of these cloud platforms only provide IaaS, while others allow the user to provision clusters with Hadoop completely installed, configured, and running. This can save a great amount of time on the management and operation of Hadoop itself. For a more comprehensive list and review the reader is referred to the May 28, 2014 Magic Quadrant for Cloud Infrastructure as a Service. Here only a few of most prominent companies will be discussed.

Given that Google was the original developer of the ideas that went on to become the basis for the open source Apache Hadoop project, their solution is an attractive choice for tech-saavy businesses looking to do Big Data analytics in the cloud. The Google Cloud Platform and Compute Engine (GCE) allow users to tap into the same infrastructure that allows Google to search billions of records in milliseconds. Google enables real-time Big Data analytics with Google BigQuery. Using this platform The Institute for Systems Biology was able to reduce the computation time for analyzing cancer data sets from 15 to 2 hours. Google features a simplified pricing model that tracks Moore's law as hardware costs have fallen and provides for sustained-use discounts. The main caution is that GCE is relatively new and consequently lacks an operational track record working with enterprise customers.

Amazon Web Services (AWS) is by far the market share leader with the most varied customer base. Elastic Map reduce (EMR) arguably provides one

of the most comprehensive cloud computing environments for Big Data analytics. The AWS EMR supports both the Amazon and MapR Hadoop distributions. The MapR M7 provides enterprise grade capabilities including high-availability, volume snapshots and mirrors. Customers should be careful to manage the AWS costs as they charge extra for business level support and some other items that are bundled with competitive offerings. AWS is beginning to face tougher competition from Google and Microsoft.

Hadoop Cluster Diagram



HDInsight is Microsoft's Hadoop on Azure implementation. Azure Infrastructure Services are relatively new and Microsoft has just begun to build an ecosystem of partners. Yet they continue to catch up and have recently reported significant query response time improvements in the 2.4 version. Microsoft also is now making Apache HBase clusters available inside Azure HDInsight. One caveat is that Azure tends to be .NET oriented with little in the way of enterprise Linux options.

The Joyent Solution for is a cloud-based hosting environment for big data projects based on Apache Hadoop. Joyent's big data servers offer bare-metal performance which is nearly 3x faster, or at 1/3 the cost, of other cloud offerings. Although they are an attractive choice for cases such as running Business Analytics and batch computing on large datasets where performance is crucial, they are considered to be geared more towards developers and have an immature partner ecosystem.

Qubole distinguishes itself from other solutions like AWS EMR by enabling orchestration of complex workflows combining Apache Hive, Pig, and map-reduce. It allows for intelligent management of the cluster costs and the customer is only charged for the computational resources that are needed. Qubole has optimized S3 support outperforming even AWS EMR and offers unique database adapters to connect popular data sources.

The advantages of using cloud based services for Hadoop include:

- Enables businesses to provision large scale resources quickly
- Simplifies some Hadoop management and operations activities
- Hardware infrastructure management efforts (i.e., staffing, training, planning, procuring hardware, hardening stand ard image builds, system deploy ment, verification testing, OS patching, system maintenance upgrades, etc.) are simplified and costs are significantly reduced.
- Good option for one-time use and computational bursts

The disadvantages include:

- Hadoop clusters typically ingest data in a predictable fashion, without the peaks and valleys that normally lend them selves to an elastic cloud deployment.
- In a public cloud there is no “rack aware ness” and there is less control over the infrastructure that is shared with other customers. This can result in degraded performance with multiple VM’s running a given physical server. It should be noted that one study demonstrated that under certaincircumstances (running 2 or 4 smaller VMs per physical machine) Hadoop on VMWare vSphere can perform up to 14% faster than on native hardware.
- Cloud storage is both slower and more expensive for data sets that just keep growing. Hadoop clusters in the cloud make the more sense when the data itself is generated in the cloud.
- Cloud solutions don’t have a well-established account base and proven track record for use with the most business critical applications

Hadoop Appliances

Intermediate between cloud-based PaaS and HaaS offerings and roll-your-own, there is the option of purchasing a Hadoop appliance. An appliance is an engineered system of hardware and software specifically designed to run Hadoop. There are appliances available running all of the leading Hadoop distributions, such as Cloudera, MapR, and Hortonworks on a range of hardware platforms. Virtually all of the major server hardware vendors including IBM, Oracle, and HP, have a Hadoop appliance. Many of these are very recent entries into the market with the Dell/Cloudera appliance just coming out being notable. Some of the appliances, like the NetApp solution for Hadoop, are specialized for use with particular hardware. If your organization is considering using Hadoop appliances, and you are already heavily invested in using hardware or database/application technology from a given vendor, it makes sense to consider their solution.

Appliance hardware is more expensive than building a comparable cluster on your own. But if the staffing, training internal IT resources, hardware build outs, building and testing the OS and Hadoop image are included, the costs become comparable to roll-your-own. The reliability and support of the bundled appliance offering can be very attractive for many organizations.

The Hadoop appliance bundles vary in the pricing models and what is included. It is important to read and understand the fine print in the contract on software pricing and upgrades. For example, IBM includes the first year of maintenance free while Oracle charges for support. The IBM and Oracle appliances are both in the range of $500,000. IBM is one of the cheaper options when the included support is considered. Oracle and Teradata install and configure the Hadoop software on-site which takes longer to get up and running. IBM ships their appliance with everything pre-installed and configured so they can be up and running in a few hours.

The IBM PureData System for Hadoop H1001 appliance, sometimes referred to as the InfoSphere BigInsights appliance, is a single-rack system that contains servers and switches, as well as the power distribution units and a KVM. The appliance contains two Hadoop master nodes in a highly available configuration. These run the appliance software which includes the IBM InfoSphere BigInsights for IBM PureData System for Hadoop application. There are 18 data nodes that hold the user data and run the BigInsights services and query processes. The appliance has 864 GB of memory and 216 TB of uncompressed disk space for user data on the appliance.

Oracle Big Data Appliance runs Oracle Linux and is based on Cloudera's Hadoop Distribution and includes Apache Hadoop with Cloudera Manager, and an open source distribution of the R data science programming language. The system also includes the Oracle NoSQL Database Community Edition and the Oracle HotSpot Java Virtual Machine. The Oracle Big Data Connectors product allows for integration of data stored in Hadoop and Oracle NoSQL Database with Oracle Database 11g. Oracle Direct Connector for Hadoop Distributed File System (ODCH) enables the Oracle Database SQL engine to access data seamlessly from the Hadoop Distributed File System. The appliance has an 18 node rack, with each node having two 8 core Intel Xeon E5-2660 processors with 64G of RAM (expandable to 512G) and 648T disk storage. The platform provides 40G/s infiniband network between the appliance nodes and from the appliance to other Oracle appliances (i.e Exadata and Exalytics). The Oracle Data Integra-

tor Application Adapter for Hadoop enables Oracle Data Integrator to generate Hadoop MapReduce programs through a graphical interface. If the integration of Oracle and Hadoop is a priority for your organization, this appliance makes great sense. Note that the connector tools are available as stand-alone utilities, so you can get good integration even if you decide to go with roll-your-own choice or another appliance.

The Netapp Hadoop Open Storage System is made of building blocks. Each block has a Netapp E-Series Storage Module and 4 data node servers that must be chosen and purchased separately by the customer. The supported software stack includes Cloudera Hadoop and Redhat Linux 5.6. In the base configuration, the rack and storage system include 4 Hadoop7 master and administration servers (including 2 name nodes and the job tracker), 1 FAS2220 storage module (optional but recommended for the master servers), 12 compute servers for data nodes, 2 Netapp E2660 (Engenio Operating System, 60x3TB disks in RAID5 or 6 configuration) storage modules that provide 360TB, and 2 Ethernet switches. An expansion rack is available that comes with an additional 16 compute servers, 4 NetApp E-Series storage modules, and 2 Ethernet switches. Hadoop was deliberately designed to use the local storage on each processing node for data locality. Moving the storage to an E2660 is a serious departure from the standard architecture. Similarly, the Hadoop HDFS redundancy features were designed to replicate each data block 3 times and the use of RAID systems is discouraged. Although the NetApp solution is a specialized hybrid that most Hadoop administrators would not be comfortable with, it is reasonable for shops that are heavily invested in NetApp technologies to consider this option.

Pivotal (formerly EMC Greenplum) offers the Pivotal Data Computing Appliance (DCA). They support the Pivotal HD Enterprise Hadoop distribution (Apache Hadoop 2.0, HDFS, and the common ecosystem tools – Pig, Hive, Zookeeper and Hbase) and the Greenplum MR distribution (based on the MapR distribution and MapRFS improvements for better performance and high availability, with support for NFS access to Hadoop's file system). Similar to the IBM and Oracle appliances, the DCA also has 18 nodes. There are 2 master servers that process the queries and come up with execution plans, and 16 segment servers (12 cores, 48 GB RAM, 28TB usable capacity) that perform the processing and store the data. The appliance comes bundled with the Greenplum relational, shared-nothing, MPP database, but the Pivotal HD Hadoop distribution module needs to be added. The EMC Isilon OneFS now provides native support for the HDFS protocol, so MapReduce jobs on a Pivotal HD cluster can work on data stored in Isilon's NAS. This solution uses the Isilon striping and redundancy algorithms instead of replicating each block 3 times as is typical in HDFS. At a glance, this combination appears to be a direct competitor to the Netapp Hadoop Open Storage System. As with the other vendors, the Pivotal DCA is an attractive option if your company uses the Greenplum database. Greenplum's relational DB integrated with Hadoop for unstructured data provide a complete platform for big-data analytics.

Dell, Cloudera and Intel have recently announced that they are collaborating on a Hadoop appliance offering designed to speed the performance of Hadoop environments by moving more data into shared memory. The Dell In-Memory Appliance for Cloudera Enterprise is the first of the Dell Engineered Systems for Cloudera Enterprise to be released. It includes the relatively new Apache Spark which is a parallel, high-speed data-analysis system for streamed data that relies on in-memory analysis. With this additional component Hadoop will enable organizations to develop

fast, unified big data applications that combine batch, streaming, and real-time, interactive business analytics. This solution uses Intel Xeon processors and ScaleMP's Versatile SMP (vSMP) architecture to aggregate multiple servers into a single virtual machine for the creation of large memory pools for in-memory processing. The pooling of memory and CPU is critical to the performance and scalability required by real-time analytics of streaming data.

The Teradata appliance runs on the SUSe Linux Enterprise Server (SleS) 11 64-bit operating system with the Hortonworks Data Platform (HDP) for Hadoop and additional Teradata software technology. The full Hadoop cabinet consists of 2 master nodes (Dual Intel eight-Core Xeon®processors @ 2.6GHz per node), 16 data nodes (Dual Intel Six-CoreXeon®processors @ 2.0GHz per node) and enterprise class storage. It can hold up to 152TB of uncompressed data per cabinet and the system can scale up to 10PB. The appliance is linked by a high throughput BYNeT V5 on a 40GB/s InfiniBand interconnect for fast data exchange between Teradata, Teradata Aster, and Hadoop systems. The software bundles the Teradata Connector for Hadoop (DCH), which is a set of APIs and tools that support high-performance, parallel, bi-directional data movement between Teradata systems and the Hadoop ecosystem software.

HP's AppSystem for Apache Hadoop offers Reference Architectures for each of the leading Hadoop distributions (Cloudera, Hortonworks, and MapR) with the HP DL380 and SL4500 server platforms. This affords the customer some flexibility in8 configuring the Hadoop cluster to best fit the organizations needs. The appliance comes with one management node (12 cores, 64GB memory, 8x900GB SAS drives, 1x1Gb Ethernet) for running the HP software stack and the Cloudera Manager, one Hadoop namenode (12 cores, 64GB memory, 4x900GB SAS drives, 1x1Gb Ethernet) one JobTracker node (12 cores, 64GB memory, 4x900GB SAS drives, 1x1Gb Ethernet), and 18 worker nodes (12 cores, 64GB memory, 16x1TB SAS drives, 1x1Gb Ethernet). HP has advertised the ability to link HDFS and Hadoop applications with their Vertica column-oriented distributed database and the Autonomy Intelligent Data Operating Layer (IDOL) 10 stack, which creates contextual information from unstructured text, audio, and video data.

The advantages of using a Hadoop Appliance include:

- Build out, OS and Hadoop patching, and cluster management are greatly simplified
- Appliances come in standard configurations
- Appliances come with the reliability of vendor support
- Eliminates learning curve for administrators on each component
- Lower set up time and enablement

The disadvantages include:

- Appliances are not as flexible as roll-your-own and have a very specific configuration with a certain number of nodes, CPU's, RAM, and storage.
- Becoming locked into a specific vendor rather than using commodity hardware as Hadoop was initially intended

Roll-your own Hadoop Cluster

After evaluating the options for running Hadoop and completing the typical proof of concept, the most ambitious, data-driven companies may choose to build and maintain their own Hadoop clusters in-house. This will entail careful capacity planning, Hadoop distribution selection, Hadoop ecosystem and data visualization tool selection, hardware selection and procurement (server, network, disk storage), development of standard OS/Software image and hardware build, cluster deployment, validation testing, and finally maintenance (patching, adding servers, replacing failed components and servers)

The questions of how much storage capacity will be need and how fast it is expected to grow need to be answered. Usually a small cluster of a hand full of nodes is deployed as an initial investment and demonstration of the technology. Often times the initial cluster is scaled out and becomes the development or production environment.

There are many options among server hardware vendors for the nodes in a Hadoop cluster. Hadoop is designed to run on commodity server hardware and large database class machines are not recommended. This is a departure from many database deployments of the past and conceptually new to many system administrators and organizations. Loss of a master server such as a name node or job tracker almost certainly means some kind of service disruption. It is recommended to use carrier class hardware for the Hadoop master servers. On the other hand, worker nodes are expected to fail regularly. Master nodes have a requirement for more RAM and CPU while data nodes tend to host more disk storage.

Another concept that is counter to the common knowledge of many storage and system administrators is the idea that RAID disk storage systems are not required or recommended for use in Hadoop clusters. HDFS handles data redundancy by replicating data blocks between cluster nodes so RAID is not needed for date protection. The Hadoop favored JBOD (Just A Bunch of Disks) configuration has also been shown to be better for performance. RAID 0 read and write operations are limited by the speed of the slowest disk in the array. Because disk performance often shows considerable variation in practice (up to tens of percent), even for disks of the same model, this can be a significant effect. In a JBOD configuration disk operations are independent, so the average speed of the round-robin read/writes to the disks is greater than that of the slowest disk.

For clusters up to about 10 nodes, there is enough capacity to run the name node and the job tracker on the same machine (as long as at least one copy of the name node's metadata is stored on a remote file system). The name node needs more memory, so the name node and job tracker should be moved onto separate machines. The secondary name node can be run on the same machine as the name node, but again for reasons of memory usage (the secondary has the same memory requirements as the primary), it is best to run it on a separate piece of hardware, especially for larger clusters.

30 to 40 servers typically fill a rack and are connected with a 1 Gb top of the rack Ethernet switch. The uplink to the core switch or router is often 10 Gb. The aggregate bandwidth between nodes on different racks is less than that between on the same rack.

Along with the hardware selection you organization must decide on the specific operating system

and Hadoop software distribution to deploy. Hadoop can be directly downloaded from the Apache web site, but in business-critical environments most companies will select a vendor supplied distribution. Vendor distributions are intended to overcome issues with the open source code and provide additional value to customers. They provide reliability by stabilizing the software and releasing timely software patches and bug fixes. They also bundle supplemental management tools and feature sets. Most importantly, they the supply technical assistance and support necessary to adopt the platforms for mission-critical and enterprise-grade tasks. Most of the vendors participate in improving the standard Apache Hadoop distribution by committing updated code to the open10 source repository. It is advisable to use an automated installation method like Red Hat Linux's Kickstart with other configuration management tools, like Puppet and Chef, for managing the installation of the same software on each node.

Although there are many flavors, the three top Hadoop distributions are provided by Cloudera, MapR, and Hortonworks. Others include IBM, Oracle, Intel, Pivotal, etc. Some of these are designed to the vendor's specific appliance implementation, for example IBM's distribution is optimized for use with the Xeon processor line. Others are available with an appliance or separate as software only. Cloudera and Hortonworks are based entirely on open source. MapR adds some proprietary components to the M3, M5, and M7 Hadoop distributions to improve the high-availability and usability in enterprise business environments. They use MapRFS file system in place of HDFS which includes improvement over the standard HDFS name node and the ability to take storage snapshots.

The advantages of using a roll-your-own Hadoop appliance include:

- Complete flexibility in choosing the best fit hardware and software stack for your business use-case

The disadvantages include:

- The need to perform the hardware builds and management with your companies own staff and resources
- Less vendor support is available than the option of using a Hadoop appliance

Conclusion

These are exciting times in the Internet age with such rich and voluminous sources of data and information. Even individuals and small businesses can benefit from the use of these technologies for business analytics and intelligence in Cloud environments. Here we have reviewed the many options for using Hadoop based clusters, from the quick and easy like HaaS in the Cloud, to the more intensive endeavors of roll-your-own. Yet it will take significant effort for your organization to sort through the offerings to determine which will provide the best business value.

Cost effective design and implement of business analytics platforms at large scales must be tightly aligned with business drivers to be successful at promoting bottom line and improving profitability. Like the past IT waves based on business application implementations for RDBMS platforms, some companies will be successful at gaining a competitive advantage over other market players, and other organizations will miss-manage the technology and fail. It is strongly recommended to complete a feasibility study and use the technology in the right context. If the analysis clearly demonstrates that your company needs to extract business-actionable information from large amounts of data in a timely manner, then Hadoop may be the platform for you.

Whether you are interested in a one-time application; or a permanent Big Data storage and computation platform as in an enterprise data hub; domain-specific vertical applications, or a horizontal use-cases across multiple industries; there is a Hadoop platform option to meet your requirements!

About Vertex

Vertex is a CMMi Level-3 IT consulting organization that engages with its customers at a strategic level and provides 'thought leadership'. Vertex's team of Solution Scientists craft innovative solutions, with a holistic view, that make businesses smarter. Vertex acts as an advisory partner, aligning its offerings with the business goals and objectives of its customers.

CMMI Microsoft Partner Silver Collaboration and Content · INFORMATICA PARTNER · TIBCO Spotfire PARTNER · NINTEX Certified PARTNER · tableau PARTNER · salesforce appexchange program partner · salesforce silver cloud alliance partner