Datavision Consulting Services:
Hadoop Network Compute Cluster

Datavision, Inc. is a leading specialty information technology consulting and staffing firm providing a full range of professional services to address core technology areas which support and drive critical business processes.

Datavision's network engineering and support services help our clients plan for Cloud-centric and OpenStack initiatives. Our consultants will work with you to assess the current environment, develop a roadmap and architecture to integrate the technology, and also help to realize the network infrastructure best suited to the business needs.

Our services include the following general stages of engagement, with individual client needs driving the complete statement of work:

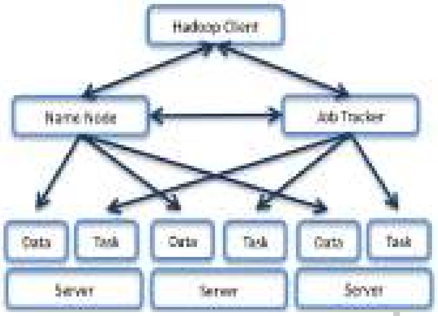
- Design & Planning
- Implementation: Integration & Migration
- Operational Support/Training
- Network Optimization

Integral to our service offer, Datavision can provide expertise to help operate the network with you on a short- or long-term basis. We can also develop training for your personnel to thoroughly understand the technology being implemented in the network, and to help ensure a smooth transition from the prior network architecture to the new.

Hadoop Overview

Hadoop and other distributed systems are increasingly the solution of choice for next generation data volumes. A high capacity, any to any, easily manageable networking layer is critical for peak Hadoop performance.

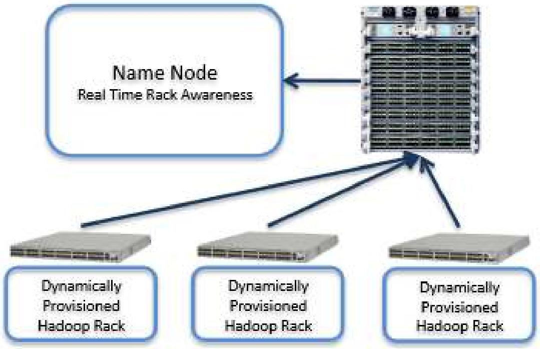
Data analytics has become a key element of the business decision process over the last decade, and the ability to process unprecedented volumes of data a consequent deliverable and differentiator in the information economy. Classic systems based on relational databases and expensive hardware, while still useful for some applications, are increasingly unattractive compared to the scalability, economics, processing power and availability offered by today's network driven distributed solutions. The perhaps most popular of these next generation systems is Hadoop, a software framework that drives the compute engines in data centers from IBM to Facebook



Hadoop and the related Hadoop Distributed File System (HDFS) form an open source framework that allows clusters of commodity hardware servers to run parallelized, data intensive workloads. Actual clusters include shoe string research analytics to thirty petabyte data warehouses, and applications range from the most advanced machine learning algorithms to distributed databases and transcoding farms. Given sufficient storage, processing, and networking resources the possibilities are nearly endless.

The HDFS stores multiple copies of data in 64MB chunks throughout the system for fault tolerance and improved availability. File location is tracked by the Hadoop **NameNode**. Replication is increased relative to frequency of use, and a number of other tunable parameters and features such as RAID can be used depending on the application. Because replication is accomplished node to node rather than top down, a well architected Hadoop cluster needs to be able to handle significant any to any traffic loads.

Parallelization and Pushing the Computation to the Data


Name Node
Real Time Rack Awareness
Dynamically Provisioned Hadoop Rack
Dynamically Provisioned Hadoop Rack
Dynamically Provisioned Hadoop Rack

The Hadoop **JobTracker** breaks down large problems into smaller computational tasks assigned to servers in the cluster. In order to handle large data sets, servers are given tasks relevant to the data already present in their directly attached storage (DAS).This is often referred to as pushing the computation to the data, and is a critical part of processing petabytes -even with 100 GbE, a badly allocated workload could take weeks to simply read in all the data necessary... Finally, **rack awareness** allows the job Tracker to assign servers close to the data in the network topology if no directly attached server is available.

Datavision Consulting Services:
Hadoop Network Compute Cluster

Hadoop and Sound Network Cluster Design

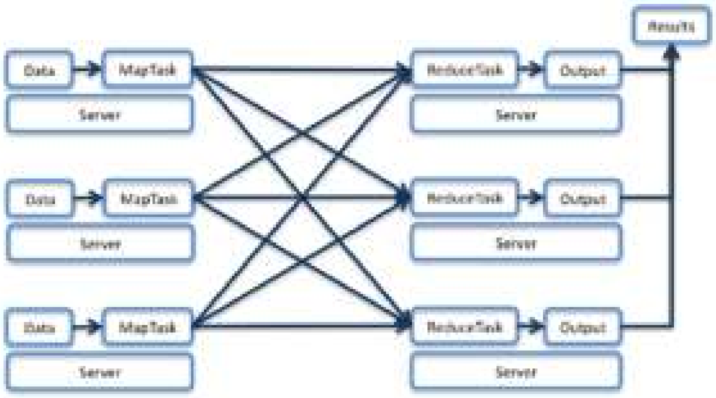
How the MapReduce Algorithm Works

MapReduce is the algorithm originally used in Google’s massively parallel web ranking systems and forms the cornerstone of the Hadoop system. It is composed of two steps: Map and Reduce.

Map: Mapping refers to the process of breaking a large file into manageable chunks that can be processed in parallel. In data warehousing applications where many types of analysis are conducted on the same data set, these chunks may have already been formed and distributed across the cluster. However, for many processes involving changing data or one time analyses, the entire multi-terabyte to multi-petabyte workload must be efficiently transferred from storage to the cluster members on a per case basis - Facebook’s larger clusters often intake 2PB per day. In these situations a high capacity network is critical to time-sensitive analytics.

Once the data has been distributed throughout the cluster, each of the servers processes the data into an intermediate format paired to a “key” which determines where it will be sent next for processing.

Reduce: When the Mapping Servers have completed their tasks, they send the intermediate data to the appropriate Reduce Server based on the data key. While many tasks have significant compression after the Mapping calculations are completed, other analyses such as the sorting used in descriptive statistics require almost the entire data set to be re-allocated, or “shuffled” to the Reduce Servers. At this point the data network is the critical path, and its performance and latency directly impact the shuffle phase of a data set reduction. High-speed, non-blocking network switches ensure that the Hadoop cluster is running at peak efficiency.



So, what’s the Point!?
Or, What is the Impact of Network Design on Hadoop Cluster Performance

Big Data/Hadoop Network Engineering

Hadoop is unique in that it has a 'rack aware' file system - it actually understands the relationship between which servers are in which cabinet and which switch supports them. With this information it is able to better distribute data and ensure that a copy of each set of data is distributed across different servers connected to different switches. This prevents any single switch failure from causing data loss. When the JobTracker distributes workload/computation to the servers that are storing data it tries to put the workload on the server co-located with the data to be mined. If that server is already being utilized then it sends the computation to another server in the same cabinet as the primary server with the data. This ensures that the network backbone is avoided for all bulk data movement except during data ingestion. The ability to moderately oversubscribe the network backbone goes up - so rather than wire-speed network bandwidth you can use a range of 3:1- 5:1, enabling a more cost effective deployment of a scale-out storage architecture.

Hadoop is also a Layer 3 aware file system - it uses IP for node addressing - this means it is routable. There is no requirement, nor benefit to building a large and flat Layer-2 network for Hadoop. You can use routing, building a scalable, stable, and easily supported ECMP network based on OSPF in the smaller deployments, and BGP in the larger ones and it will be stable and contain broadcasts and faults to each cabinet.

Hadoop Big Data Network Engineering

Operationally these are simpler networks to troubleshoot and maintain with full toolsets available in every host stack and network element: Traceroute, Ping, Arping, fping, etc are available for L3 network day-to-day troubleshooting without requiring customer tool development or locking yourself into a single-vendor proprietary architecture.

Redundancy: Two switches at the top of each cabinet is a common enterprise recommendation - it ensures that any switch failure doesn't bring any server down. In a traditional enterprise data center it is a common recommendation. However in a Hadoop cluster customers have a choice - something every other vendor, will not usually recommend: go with a single ToR switch for all cabinets except for the main cabinet that keeps the NameNode and JobTracker servers.

If the NameNode and/or JobTracker fail the job stops and the cluster fails, these two servers need to be well protected. However, once you exceed ten cabinets any single switch failure will reduce processing capacity by less than 10% and a declining percentage as the cluster scales out. The decision to use network redundancy at the leaf/access-layer becomes an operational decision around network upgrade process more than it becomes about data integrity and data availability as we increasingly trust the file-system and application tiers to handle this responsibility.

Data Integrity and Optimization:

Because Hadoop is network-aware and the organization of the data structures is based, in part, on the network topology ensuring that we have an accurate mapping of network topology to servers is of paramount importance. (As an example, with Arista EOS clients can load extensions that let them automatically update the Hadoop Rack Awareness configuration.

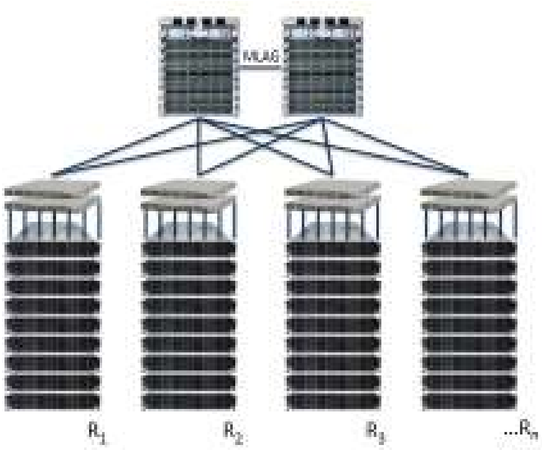
This ensures several important things:

1) Data is distributed properly across servers so no single point of failure exists. Misconfiguration, or a lack of configuration, could inadvertently enable the NameNode to 'distribute' the data to three separate storage nodes that are all connected to the same switch. A switch failure then causes data loss and the data mining job stops or worse has invalid results.

2) As the Hadoop cluster scales no human has to maintain the Rack Awareness section, it is automatically updated.

3) As nodes age and are replaced or upgraded the topology self-constructs and data is automatically distributed properly.

4) Performance is improved and deterministic because no data is ever more than one network 'hop' (single MAC/IP lookup, no proprietary fabric semantics or tunneling games) from the computation that depends on that data. Jobs get distributed to the right place.



Datavision Consulting Services:
Hadoop Network Compute Cluster

datavision
strategy / design / realization

Hadoop Big Data Network Engineering

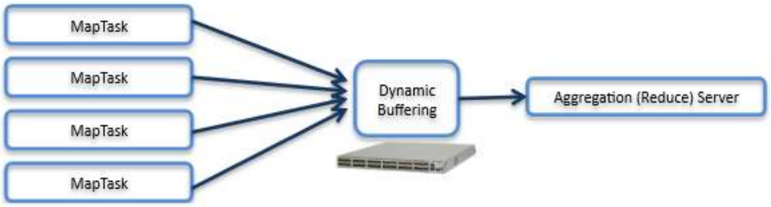
Performance: Hadoop performs best with a wire-speed Rack switch. This helps with data ingestion which is the largest bulk data move the network has to absorb because of the Hadoop Rack Awareness architecture, but more importantly during all operational runtime it eliminates worry and simplifies trouble-shooting. If the switch is wirespeed you only have to worry about the uplinks being congested, not the switch fabric itself causing drops. (and simplified troubleshooting.) The amount of network capacity between the leaf and spine can be decreased below 1:1 because of the Rack Awareness. This enables us to build larger networks for less cost

A network designed for Hadoop applications, rather than standard enterprise applications, can make a significant difference in the performance of the cluster.

High Capacity, Any-to-Any Topology, and Incremental Scalability: Getting data into the cluster can be the first bottleneck, and whether replicating or shuffling data, Hadoop requires significant any-to-any node traffic to get its job done. In order to efficiently access stored results or simply calculate new ones, a well-provisioned network with full any-to-any capacity, low latency, and high bandwidth can significantly improve Hadoop cluster performance. And, as workloads grow, it is important that the network can sustain the inclusion of additional servers in an incremental fashion - Hadoop only scales in proportion to the compute resources networked together at any one time.

High Availability and Fault Tolerance: Even though Hadoop has self-contained fault tolerance in any single node, a failure in network connectivity in any but the largest clusters will stop HDFS data accessibility and any currently running jobs. Highly available, fault-tolerant networking equipment and architecture can make sure that the Hadoop cluster stays and assist in a quick re-provisioning of any failed server nodes. Datavision can help you architect, design and deploy the optimal network con-

Dynamic Buffers and Visibility



Traffic fan-in is an unavoidable fact of aggregation. Networks employing dynamically allocated buffers can shift resources to congested ports in real time for superior adaptability in the face of rapidly changing traffic workloads. If dynamically allocated network buffers are employed, even the most oversubscribed **Reduce** server can receive all its intermediate data without lost packets and the consequent network overload and inefficiency that would otherwise occur. Finally, tools such as Arista's Latency Analyzer allow network introspection and cluster reconfiguration to eliminate bottlenecks and create workload dependent cluster optimization.

Management And Extensibility

Getting the most out of any scaled solution requires proper management tools and a framework for customized application needs. To use the example of EOS, because it is based on Linux, EOS provides the perfect foundation for leveraging open source tools and creating user defined functionality Admins gain immediate productivity with their preferred binaries and scripts without needing to learn and relearn proprietary operating systems. Cluster management systems such as perfSONAR, gmond, Nagios or Ganglia can be run directly on the switch to detect and proactively address any unexpected data center conditions-possible responses range from email and SMS alerts to actions immediately shifting topology and configuration. EOS

At Datavision, your success is our business.