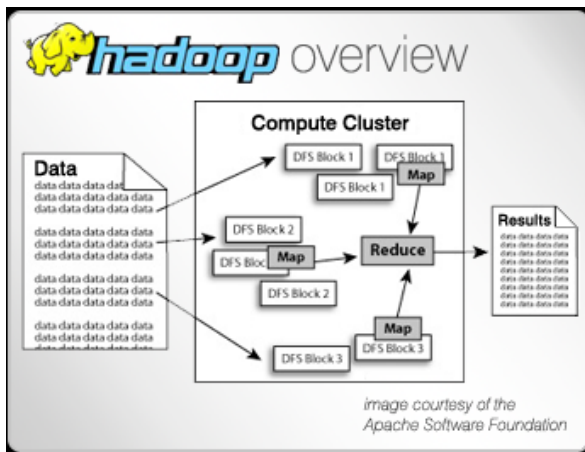


Big Data Imperative - Hadoop

Hadoop Introduction

An increasingly complex world combined with the vast proliferation of data and the burning need to stay two steps ahead of the competition has amplified focus on leveraging analytics within organizations. The amount of data in our world is exploding, and analyzing large data sets, otherwise known as Big Data, is quickly becoming a key foundation of competition, laying the groundwork for new waves of productivity, growth, and innovation.

A highly scalable, fault-tolerant distributed system for data storage is the core technology necessary to enable this level of growth and productivity. The scalability is made possible with a Self-Healing High Bandwidth Clustered Storage composed of HDFS (Hadoop Distributed File System) and a fault-tolerant Distributed Processing framework known as MapReduce.



Hadoop runs on a collection of commodity servers and is an ideal platform for consolidating large-scale data from a variety of new and legacy sources. It complements existing data management solutions with new analytical and processing tools. It delivers direct value to companies in an array of vertical markets such as Social Media & Web/Digital Services, Health & Life Sciences, Financial Services, Government & Research, and Telecommunications.

To efficiently and effectively manage, use and access the increasingly large petabytes of data, organizations are in need of an established and reliable Big Data storage partner.

The Big Data Impact – How Does It Work?

According to market research firm IDC's Digital Universe publication, 1.8 trillion gigabytes of data will be created and replicated in 2011 and growing fast (it has grown by a factor of 9 in just five years). This is equivalent to the information contained on 250 billion DVDs. With the amount of data in the world increasing at exponential rates, analyzing that data and producing intelligence from it becomes very important.

Hadoop is the fault-tolerant, efficient mechanism for sequential data analysis and provides a highly scalable distributed file system, which is used for storing, managing, and securing the data, while the MapReduce parallel batch processing framework provides a powerful programming model capable of harnessing the computing power of several commodity servers into a single high-performance computing cluster capable of efficiently analyzing enormous datasets.

Hadoop deployments can have very large infrastructure requirements, and hardware and software choices made at design time can have a considerable impact on the performance of Hadoop clusters. Hadoop cluster performance and ROI are highly dependent on CPU, storage, and network technology choices. In many cases, a Hadoop job can encounter bottlenecks processing data (CPU-bound) or reading data from the network or the disk (I/O-bound). Those looking to build a Hadoop cluster that don't yet understand their workload often find that the first jobs that they run with Hadoop are far different than the jobs that are processed once committed to a production environment.



For this reason it makes sense to invest in a Hadoop cluster that is balanced in CPU, network and disk I/O performance when unfamiliar with the types of jobs being run. Balancing the performance of server network I/O may improve the efficiency of every server in the cluster, thus improving the performance of the entire Hadoop infrastructure. However, without a balanced CPU and storage setup with performance to match, Hadoop cluster performance can still be limited.

AMAX Engineers have created a highly scalable, balanced Hadoop 20-node cluster configuration to deliver performance without limitation.

Each node is based on the AMAX ServMax SX-2208, and key features include:

2 x NameNodes and 18 x DataNodes with:

- 2U Server w/ dual Intel® Xeon® 5645, 2.40Ghz, 6-Core processors
- 48GB(NameNode) / 24GB(DataNode), DDR3, 1333Mhz ECC Registered
- 6 x 600GB 6Gb/s SAS(NameNode) / 12 x 2TB 6Gb/s SAS(DataNode)
- 9211-8i 6Gb/s SAS HBA
- Emulex OneConnect OCe11102-NX Dual-port 10GbE Adapter
- 1x 24 port 10GbE switch
- 1x 24 port GbE switch

A balanced Hadoop configuration relies heavily on specific application-tuning for the best performance, and AMAX offers additional Professional Services in the form of Hadoop Engineering expertise to further fine-tune a Hadoop cluster for specific applications. AMAX's team of Professional Services Engineers have many years of experience and abundant industry certifications in all facets of network and infrastructure design, integration, analysis and monitoring. AMAX can work closely with you to ensure the infrastructure you install today will deliver the performance you need to stay competitive tomorrow.

Scientific applications can also take advantage of Hadoop and its transparent data replication, data locality aware scheduling, and fault tolerance capabilities. Video and image analysis, log processing and data warehousing are all areas in which Hadoop is currently deployed.



Hadoop Use Cases

Use Case	Application	Industry	Application	Use Case
ADVANCED ANALYTICS	Social Network Analysis	Web	Clickstream Sessionization	DATA PROCESSING
	Content Optimization	Media	Clickstream Sessionization	
	Network Analytics	Telco	Mediation	
	Loyalty & Promotions Analysis	Retail	Data Factory	
	Fraud Analysis	Financial	Trade Reconciliation	
	Entity Analysis	Federal	SIGINT	
	Sequencing Analysis	Bioinformatics	Genome Mapping	

About AMAX

For more than 30 years, AMAX has provided customers with dynamic computing solutions utilizing the latest and most advanced IT technologies. By leveraging countless years of IT expertise and dedication to an open architecture design, AMAX excels at delivering computing solutions that offer exceptional efficiency without sacrificing power or performance, enabling highly optimized and scalable computing environments designed to maximize server utilization to workload and still provide ample headroom for growth. AMAX's award-winning, industry-leading engineering and manufacturing expertise enable us to deliver a comprehensive line of computing solutions, including innovative HPC servers and clusters to enterprise-class storage solutions, supporting a wide range of industries. Our solutions are fully optimized for virtualized computing environments, sophisticated database and datacenter usage, enterprise and high performance computing, web/cloud, as well as intricate IT infrastructures and collaboration applications.

Industry Reliance on Big Data Analysis

Netflix is a service which offers online DVD rental-by-mail and video streaming in the United States. It has over 100,000 titles and 10 million subscribers, over 55 million DVDs and, on average, ships 1.9 million DVDs to customers around the world each day. Netflix offers video streaming services, enabling the viewing of films directly on a PC, TV or mobile device. At the heart of their movie recommendation algorithm is Hadoop, HDFS, and MapReduce, which is used for query processing and Business Intelligence. Like Netflix, multi-billion dollar corporations such as Yahoo, Facebook, Microsoft, Amazon and many others also rely on Hadoop to run large distributed computations.



For more information about AMAX solutions for Big Data, please visit www.amax.com or send your inquiry to sales@amax.com and an AMAX solutions architect will further assist you.

