



**A Primer**  
by  
**Steve Staso**  
**Sun Microsystems, Inc.**  
**Last edited: July 10, 2008**

# What is Hadoop

- Hadoop is a distributed computing platform written in Java. It incorporates features similar to those of the Google File System and of MapReduce to process vast amounts of data
  - > Current release is 0.17.1
  - > “Hadoop is a Free Java software framework that supports data intensive distributed applications running on large clusters of commodity computers. It enables applications to easily scale out to thousands of nodes and petabytes of data” (Wikipedia)
- What platform does Hadoop run on?
  - > Java 1.5.x or higher, preferably from Sun
  - > Linux; Windows for development; Solaris, but not documented

# Hadoop is Especially Useful

- Scalable: Hadoop can reliably store and process petabytes.
- Economical: It distributes the data and processing across clusters of commonly available computers. These clusters can number into the thousands of nodes.
- Efficient: By distributing the data, Hadoop can process it in parallel on the nodes where the data is located. This makes it extremely rapid.
- Reliable: Hadoop automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures.

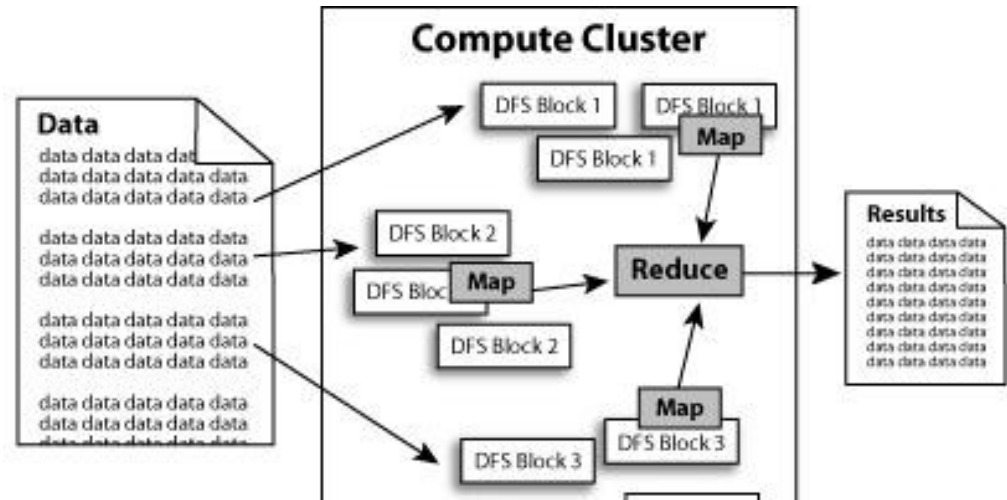
# Hadoop History

- Dec 2004 – Google GFS paper published
- July 2005 – Nutch (another Apache project - a web site crawler/search engine) uses MapReduce
- Jan 2006 – Doug Cutting joins Yahoo!
  - > Hadoop is Named after Doug Cutting's child's stuffed elephant
- Feb 2006 – Becomes Lucene subproject
- Apr 2007 – Yahoo! on 1000-node cluster
- Jan 2008 – An Apache Top Level Project
- Feb 2008 – Yahoo! Uses it in production for search index (replaced their proprietary algorithm)

# Hadoop Components

- Hadoop Distributed File System - HDFS

- > Java interface
- > Shell interface
- > C - libhdfs
- > HTTP FileSystem
- > Web interface
- > Eclipse plugin



- Hadoop MapReduce

- > Java api
- > Streaming api – via stdin/stdout
- > Pipes C++ api - via sockets

- Hadoop on Demand - tools to provision and manage dynamic setup and teardown of Hadoop nodes

# Hadoop Distributed File System

- Single Namespace for entire cluster
  - > Data Coherency
  - > Write-once-read-many access model
- Files are broken up into blocks
  - > Typically 128 MB block size
  - > Each block replicated on multiple DataNodes - default is 3x (2 in same rack, 1 in another)
- Intelligent Client
  - > Client can find location of blocks
  - > Client accesses data directly from DataNode
- Assumptions:
  - > Expects HW failures
  - > Streaming Data Access vs low Latency
  - > Large Data Sets
  - > Simple Coherency Model
  - > Moving Computation is Cheaper than Moving Data
- Not Posix compliant - still needs an underlying filesystem

# What is MapReduce?

- A Common design pattern in data processing  
cat \* | grep | sort | unique -c | cat > file  
input | **map** | shuffle | **reduce** | output
- Operates on key/value pairs
  - > Mapper: Given a line of text, break it into words and output the word and the count of 1:
    - “hi Eagle bye Eagle” -> (“hi”, 1), (“Eagle”, 1), (“bye”, 1), (“Eagle”, 1)
  - > Combiner/Reducer: Given a word and a set of counts, output the word and the sum
    - (“Eagle”, [1, 1]) -> (“Eagle”, 2)
- Used for:
  - > Log processing
  - > Web search indexing and ranking (think Google)
  - > Ad-hoc queries
- Hadoop MR has Near Linear Scalability to 2000 nodes

# Hadoop Subprojects

- Pig (Initiated by Yahoo!)
  - > High-level language for data analysis
- HBase (initiated by Powerset)
  - > Table storage for semi-structured data
  - > Modelled on Google's Bigtable
  - > Row/column store
  - > Billions of rows x millions of columns
  - > Column-oriented – nulls are free
  - > Untyped – stores byte[]
- Zookeeper (Initiated by Yahoo!)
  - > Coordinating distributed applications
- Hive (initiated by Facebook, coming soon)
  - > SQL-like Query language and Metastore
- Mahout = means Elephant Driver
  - > Machine learning - pre-built algorithms

# Who is Using Hadoop?

- Yahoo - 10,000+ nodes, multi PB of storage
- Google
- Facebook - 320 nodes, 1.3PB - reporting, analytics
- MySpace - 80 nodes, 80TB - analytics
- Hadoop Korean user group - 50 nodes
- IBM - in conjunction with universities
- Joost - analytics
- Koubai.com - community and search, China - analytics
- Last.fm - 55 nodes, 85TB - chart calculation, analytics
- New York Times - image conversion
- Powerset - Search - 400 instances on EC2, S3
- Veoh - analytics, search, recommendations
- Rackspace - processing "several hundred gigabytes of email log data" every day

# Hadoop - Typical Node configuration

- 2P 2C CPU's
- 4-8GB; ECC preferred, though more expensive
- 2 x 250GB SATA drives
- Cost is about \$2-5K
- 1-5 TB external storage

# Hadoop in the Cloud

from Wikipedia

- Hadoop on Amazon EC2/S3 services

- > It's possible to run Hadoop on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3)[9]. As an example The New York Times used 100 Amazon EC2 instances and a Hadoop application to process 4TB of raw image TIFF data (stored in S3) into 1.1 million finished PDFs in the space of 24 hours at a computation cost of just \$240[10].
- > There is support for the S3 filesystem in Hadoop distributions, and the Hadoop team generates EC2 machine images after every release. From a pure performance perspective, Hadoop on S3/EC2 is inefficient, as the S3 filesystem is remote and delays returning from every write operation until the data is guaranteed to not be lost. This removes the locality advantages of Hadoop, which schedules work near data to save on network load. However, as Hadoop-on-EC2 is the primary mass-market way to run Hadoop without one's own private cluster, the performance detail is clearly felt to be acceptable to the users.

- Hadoop with Sun Grid Engine

- > Hadoop can also be used in compute farms and high-performance computing environments. Integration with Sun Grid Engine was released, and running Hadoop on Sun Grid is possible. [11] Note that, as with EC2/S3, the CPU-time scheduler appears to be unaware of the locality of the data. A key feature of the Hadoop Runtime, "do the work in the same server or rack as the data" is therefore lost.

# Hadoop Resources

- <http://hadoop.apache.org/core/>
- <http://en.wikipedia.org/wiki/Hadoop>
- <http://wiki.apache.org/hadoop/ProjectDescription>