Hadoop and Hadoop Distributed File System (HDFS)

趨勢科技研發實驗室

Public 2009/5/13

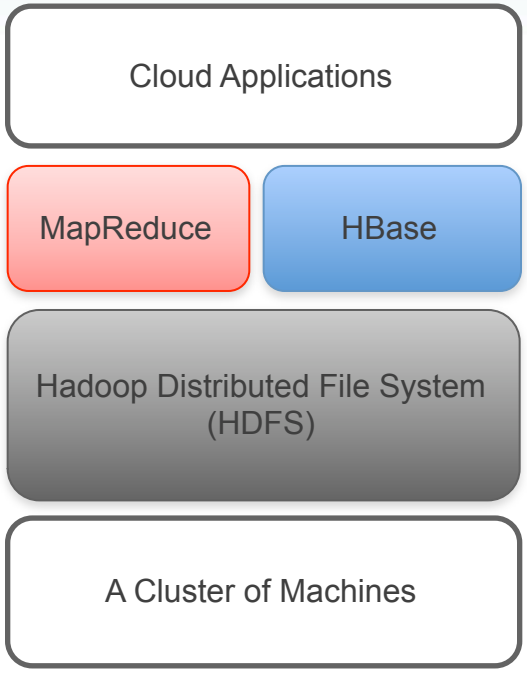
Outline

- Introduction to **Hadoop** project
- Introduction to **Hadoop Distributed File System**
  - Architecture
  - Administration
  - Client Interface
- Reference

Copyright 2009 - Trend Micro Inc.

What is Hadoop?

- Hadoop is a cloud computing platform for **processing and keeping vast amount of data.**
- Apache top-level project
- Open Source
- Hadoop Core includes
  - **Hadoop Distributed File System** (HDFS)
  - **MapReduce** framework
- Hadoop subprojects
  - **HBase**, Zookeeper,…
- Written in **Java**.
- Client interfaces come in **C++/Java/Shell Scripting**
- Runs on
  - Linux, Mac OS/X, Windows, and Solaris
  - **Commodity hardware**



Copyright 2009 - Trend Micro Inc.

A Brief History of Hadoop

- 2003.2
  - First MapReduce library written at Google
- 2003.10
  - Google File System paper published
- 2004.12
  - Google MapReduce paper published
- 2005.7
  - Doug Cutting reports that Nutch now use new MapReduce implementation
- 2006.2
  - Hadoop code moves out of Nutch into new Lucene sub-project
- 2006.11
  - Google Bigtable paper published

Copyright 2009 - Trend Micro Inc.

A Brief History of Hadoop (continue)

- 2007.2
  - First HBase code drop from Mike Cafarella
- 2007.4
  - Yahoo! Running Hadoop on 1000-node cluster
- 2008.1
  - Hadoop made an Apache Top Level Project

Copyright 2009 - Trend Micro Inc.

Who use Hadoop?

- **Yahoo!**
  - More than 100,000 CPUs in ~20,000 computers running Hadoop
- **Google**
  - University Initiative to Address Internet-Scale Computing Challenges
- **Amazon**
  - Amazon builds product search indices using the streaming API and pre-existing C++, Perl, and Python tools.
  - Process millions of sessions daily for analytics
- **IBM**
  - Blue Cloud Computing Clusters
- **Trend Micro**
  - Threat solution research
- More…
  - http://wiki.apache.org/hadoop/PoweredBy

Copyright 2009 - Trend Micro Inc.

Introduction to

Hadoop Distributed File System

Copyright 2009 - Trend Micro Inc.

HDFS Design

- **Single Namespace for entire cluster**
- **Very Large Distributed File System**
  - 10K nodes, 100 million files, 10 PB
- **Data Coherency**
  - Write-once-read-many access model
  - A file once created, written and closed need not be changed.
  - Appending write to existing files (in the future)
- **Files are broken up into blocks**
  - Typically 128 MB block size
  - Each block replicated on multiple DataNodes

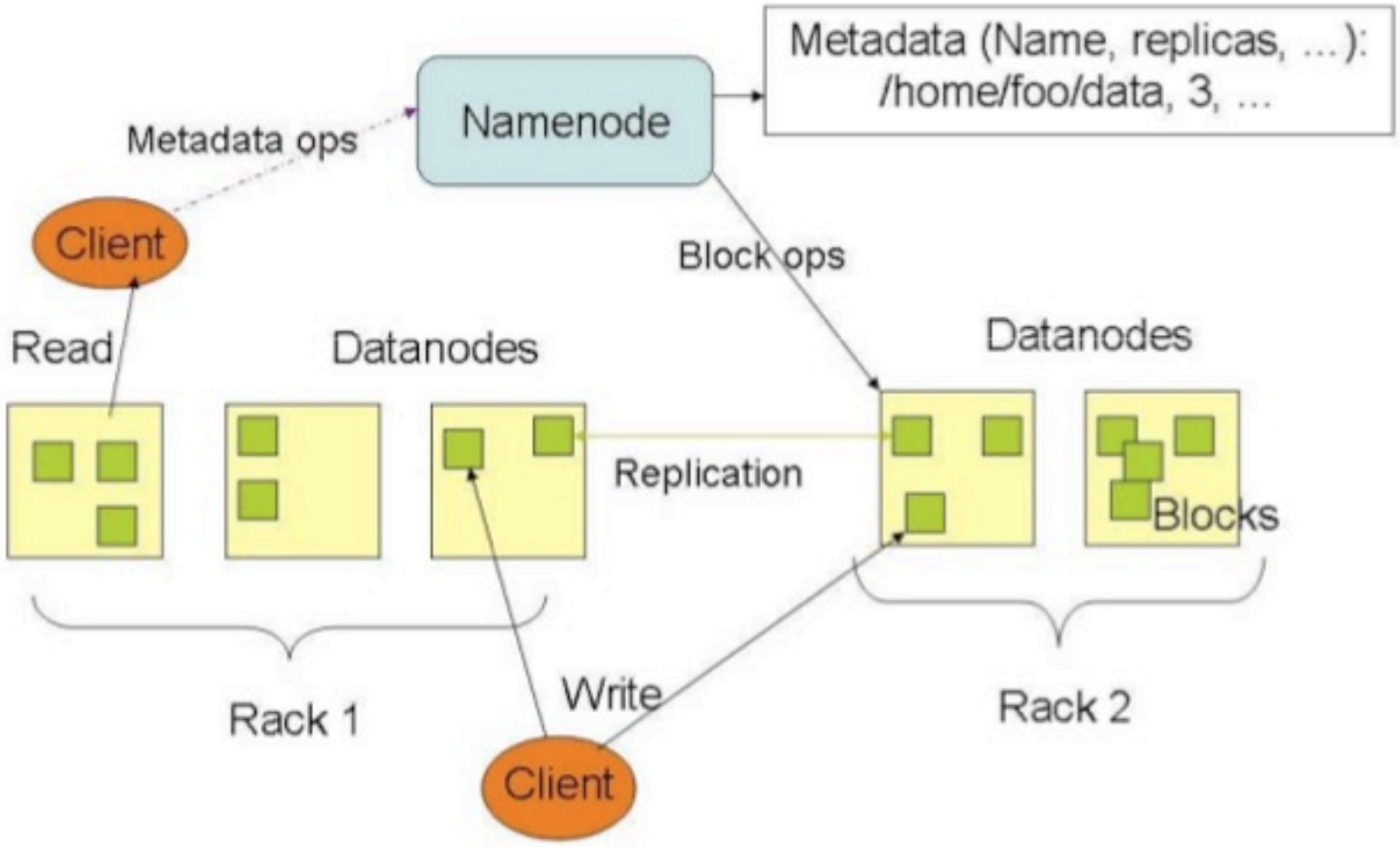
Copyright 2009 - Trend Micro Inc.

HDFS Design

- **Moving computation is cheaper than moving data**
  - Data locations exposed so that computations can move to where data resides
- **File replication**
  - Default is 3 copies.
  - Configurable by clients
- **Assumes Commodity Hardware**
  - Files are replicated to handle hardware failure
  - Detect failures and recovers from them
- **Streaming Data Access**
  - High throughput of data access rather than low latency of data access
  - Optimized for Batch Processing

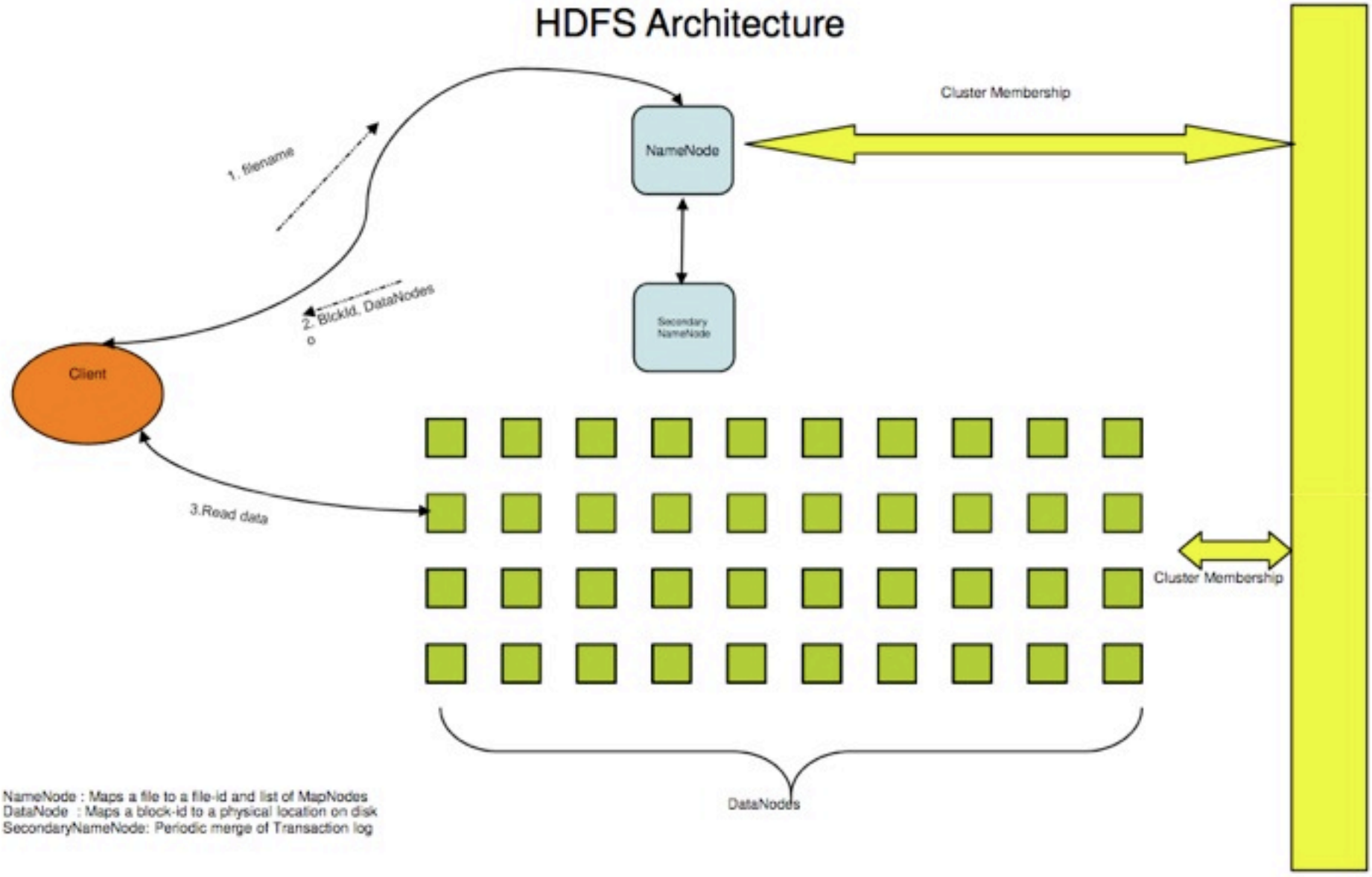
Copyright 2009 - Trend Micro Inc.

HDFS Architecture



Copyright 2009 - Trend Micro Inc.

HDFS Architecture



NameNode : Maps a file to a file-id and list of MapNodes
DataNode : Maps a block-id to a physical location on disk
SecondaryNameNode: Periodic merge of Transaction log

Copyright 2009 - Trend Micro Inc.

NameNode

- **Manages File System Namespace**
  - Maps a file name to a set of blocks
  - Maps a block to the DataNodes where it resides
- **Cluster Configuration Management**
- **Replication Engine for Blocks**

Copyright 2009 - Trend Micro Inc.

NameNode Metadata

- **Meta-data in Memory**
  - The entire metadata is in main memory
  - No demand paging of meta-data
- **Types of Metadata**
  - List of files
  - List of Blocks for each file
  - List of DataNodes for each block
  - File attributes, e.g creation time, replication factor

Copyright 2009 - Trend Micro Inc.

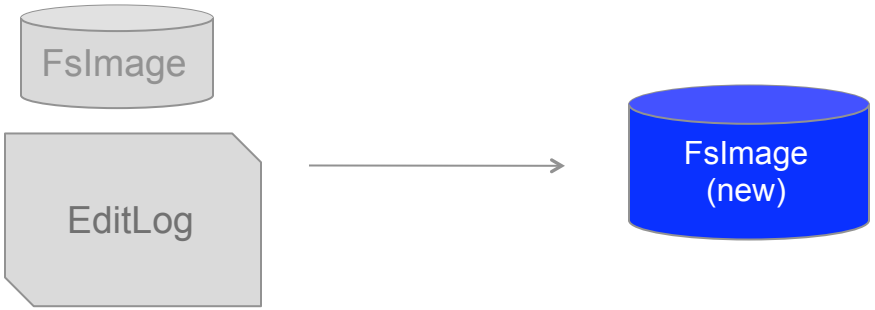
NameNode Metadata (cont.)

- **A Transaction Log (called *EditLog*)**
  - Records file creations, file deletions. Etc
- ***FsImage***
  - The entire namespace, mapping of blocks to files and file system properties are stored in a file called ***FsImage***.
  - NameNode can be configured to maintain multiple copies of *FsImage* and *EditLog*.
- **Checkpoint**
  - Occur when NameNode startup
  - Read *FsImange* and *EditLog* from disk and apply all transactions from the *EditLog* to the in-memory representation of the *FsImange*
  - Then flushes out the new version into new *FsImage*

Copyright 2009 - Trend Micro Inc.

Secondary NameNode

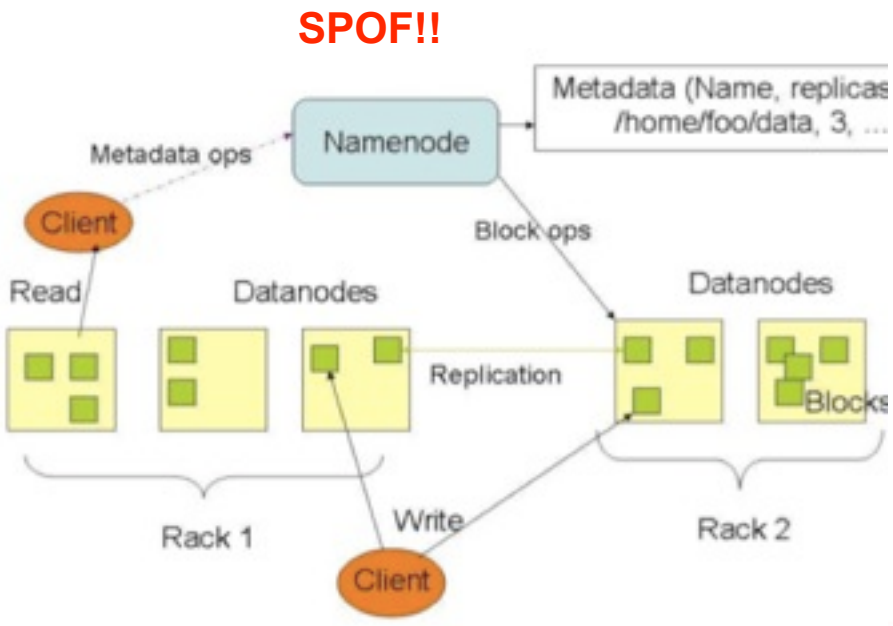
- Copies *FsImage* and *EditLog* from NameNode to a temporary directory
- Merges FSImage and *EditLog* into a new *FSImage* in temporary directory.
- Uploads new *FSImage* to the NameNode
  - Transaction Log on NameNode is purged



Copyright 2009 - Trend Micro Inc.

NameNode Failure

- **A single point of failure**
- **Transaction Log stored in multiple directories**
  - A directory on the local file system
  - A directory on a remote file system (NFS/CIFS)
- **Need to develop a real HA solution**



Copyright 2009 - Trend Micro Inc.

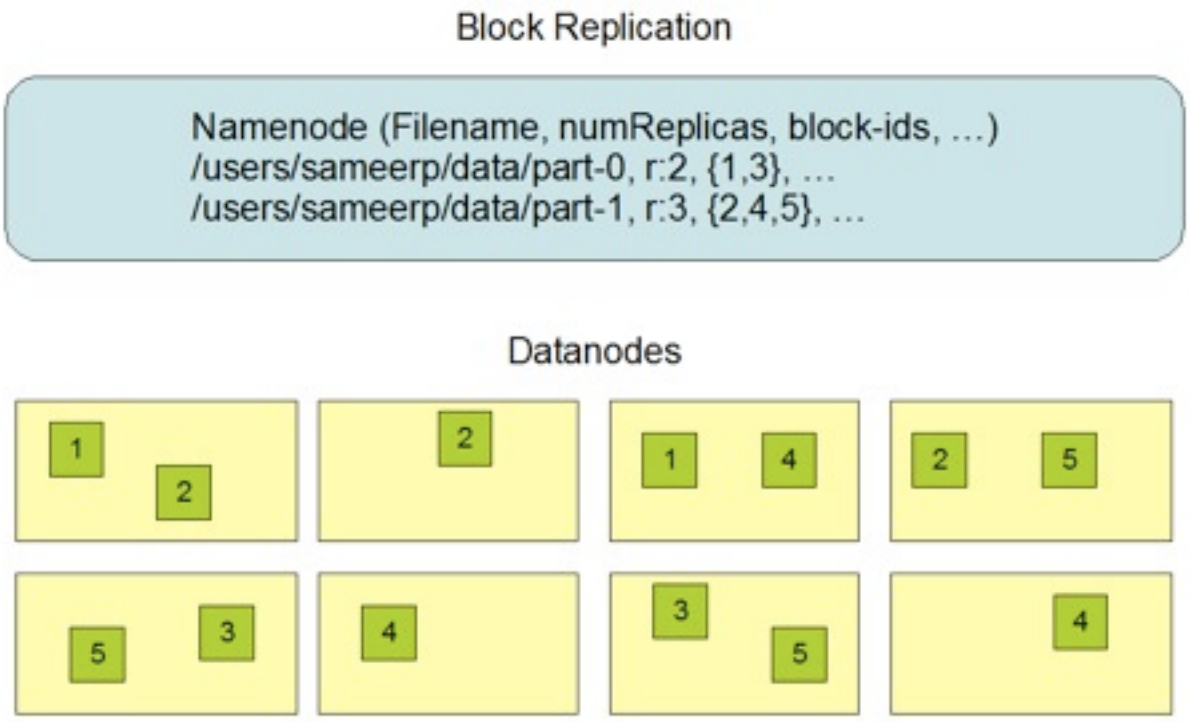
DataNode

- **A Block Server**
  - Stores data in the local file system (e.g. ext3)
  - Stores meta-data of a block (e.g. CRC)
  - Serves data and meta-data to Clients
  - Serves data and meta-data to Clients
- **Block Report**
  - Periodically sends a report of all existing blocks to the NameNode
- **Facilitates Pipelining of Data**
  - Forwards data to other specified DataNodes

Copyright 2009 - Trend Micro Inc.

HDFS - Replication

- Default is **3x** replication.
- The block size and ***replication factor*** are configured per file.
- Block placement algorithm is ***rack-aware***



Copyright 2009 - Trend Micro Inc.

Block Placement

- **Strategy (v0.19. doc, 2009.2.25, in progress)**
  - One replica on local node in the local rack
  - Second replica on a different node in the local rack
  - Third replica in the remote rack
  - Additional replicas are randomly placed
- **Clients read from nearest replica**

Copyright 2009 - Trend Micro Inc.

Heartbeats

- **DataNodes send heartbeat to the NameNode**
  - Once every ***3 seconds***
- **NameNode used heartbeats to detect DataNode failure**

Copyright 2009 - Trend Micro Inc.

Data Correctness

- **Use Checksums to validate data**
  - Use ***CRC32***
- **File Creation**
  - Client computes checksum per 512 byte
  - DataNode stores the checksum
- **File access**
  - Client retrieves the data and checksum from
  - If Validation fails, Client tries other replicas

Copyright 2009 - Trend Micro Inc.

User Interface

- API
  - Java API
  - C language wrapper (*libhdfs*) for the Java API is also available
- POSIX like command
  - hadoop dfs -mkdir /foodir
  - hadoop dfs -cat /foodir/myfile.txt
  - hadoop dfs -rm /foodir myfile.txt
- HDFS Admin
  - bin/hadoop dfsadmin –safemode
  - bin/hadoop dfsadmin –report
  - bin/hadoop dfsadmin -refreshNodes
- Web Interface
  - Ex: http://172.16.203.136:50070

Copyright 2009 - Trend Micro Inc.

Web Interface
(http://172.16.203.136:50070)

NameNode '172.16.203.136:8020'

Started:	Sun May 17 11:52:41 CST 2009
Version:	0.20.0, r763504
Compiled:	Thu Apr 9 05:18:40 UTC 2009 by ndaley
Upgrades:	There are no upgrades in progress.

Browse the filesystem
Namenode Logs

Cluster Summary

55 files and directories, 52 blocks = 107 total. Heap Size is 4.94 MB / 198.5 MB (2%)

Configured Capacity	:	13.32 GB
DFS Used	:	881.47 KB
Non DFS Used	:	6.88 GB
DFS Remaining	:	6.44 GB
DFS Used%	:	0.01 %
DFS Remaining%	:	48.35 %
Live Nodes	:	2
Dead Nodes	:	0

Live Datanodes : 2

Node	Last Contact	Admin State	Configured Capacity (GB)	Used (GB)	Non DFS Used (GB)	Remaining (GB)	Used (%)	Used (%)
172.16.203.132	2	In Service	6.66	0	3.06	3.6	0.01	
172.16.203.133	2	In Service	6.66	0	3.82	2.84	0.01	

NameNode Storage:

Storage Directory	Type	State
/tmp/namenode	IMAGE_AND_EDITS	Active

Web Interface
(http://172.16.203.136:50070)
Browse the file system

Contents of directory /test

Goto : /test go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
input	dir				2009-05-16 15:15	rwxr-xr-x	root	supergroup
ouput	dir				2009-05-16 15:17	rwxr-xr-x	root	supergroup
output	dir				2009-05-16 15:15	rwxr-xr-x	root	supergroup
output2	dir				2009-05-16 15:25	rwxr-xr-x	root	supergroup
output5	dir				2009-05-17 11:58	rwxr-xr-x	root	supergroup
output6	dir				2009-05-17 12:02	rwxr-xr-x	root	supergroup

Contents of directory /test/output6

Goto : /test/output6 go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_logs	dir				2009-05-17 12:02	rwxr-xr-x	root	supergroup
part-00000	file	0.01 KB	3	64 MB	2009-05-17 12:02	rw-r--r--	root	supergroup

Copyright 2009 - Trend Micro Inc.

POSIX Like command

```
hadoop fs [-fs <local | file system URI>] [-conf <configuration file>]
    [-D <property=value>] [-ls <path>] [-lsr <path>] [-du <path>]
    [-dus <path>] [-mv <src> <dst>] [-cp <src> <dst>] [-rm <src>]
    [-rmr <src>] [-put <localsrc> <dst>] [-copyFromLocal <localsrc> <dst>]
    [-moveFromLocal <localsrc> <dst>] [-get <src> <localdst>]
    [-getmerge <src> <localdst> [addnl]] [-cat <src>]
    [-copyToLocal <src><localdst>] [-moveToLocal <src> <localdst>]
    [-mkdir <path>] [-report] [-setrep [-R] [-w] <rep> <path/file>]
    [-touchz <path>] [-test -[ezd] <path>] [-stat [format] <path>]
    [-tail [-f] <path>] [-text <path>]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-chgrp [-R] GROUP PATH...]
    [-help [cmd]]
```

```
[root@172 bin]# ./hadoop dfs -mkdir /test/input10
[root@172 bin]# ./hadoop dfs -put ../conf/* /test/input10/
[root@172 bin]# ./hadoop dfs -ls /test/input10/
Found 13 items
-rw-r--r--   3 root supergroup       6275 2009-05-17 12:21 /test/input10/capacity-scheduler.xml
-rw-r--r--   3 root supergroup        535 2009-05-17 12:21 /test/input10/configuration.xsl
-rw-r--r--   3 root supergroup        270 2009-05-17 12:21 /test/input10/core-site.xml
-rw-r--r--   3 root supergroup       2296 2009-05-17 12:21 /test/input10/hadoop-env.sh
-rw-r--r--   3 root supergroup       1245 2009-05-17 12:21 /test/input10/hadoop-metrics.properties
-rw-r--r--   3 root supergroup       4190 2009-05-17 12:21 /test/input10/hadoop-policy.xml
-rw-r--r--   3 root supergroup        259 2009-05-17 12:21 /test/input10/hdfs-site.xml
-rw-r--r--   3 root supergroup       2815 2009-05-17 12:21 /test/input10/log4j.properties
-rw-r--r--   3 root supergroup        272 2009-05-17 12:21 /test/input10/mapred-site.xml
-rw-r--r--   3 root supergroup         10 2009-05-17 12:21 /test/input10/masters
-rw-r--r--   3 root supergroup         30 2009-05-17 12:21 /test/input10/slaves
-rw-r--r--   3 root supergroup       1243 2009-05-17 12:21 /test/input10/ssl-client.xml.example
-rw-r--r--   3 root supergroup       1195 2009-05-17 12:21 /test/input10/ssl-server.xml.example
[root@172 bin]# ./hadoop dfs -tail /test/input10/masters
localhost
[root@172 bin]#
```

pyright 2009 - Trend Micro Inc.

Java API

- Latest API
  - http://hadoop.apache.org/core/docs/current/api/

```
URI uri = new URI("hdfs://namenode/");
FileSystem fs = FileSystem.get(uri, new Configuration());
Path file = new Path("answer");

DataOutputStream out = fs.create(file);
out.writeInt(42);
out.close();

DataInputStream in = fs.open(file);
System.out.println(in.readInt());
in.close();

fs.delete(file);
```

Copyright 2009 - Trend Micro Inc.

Reference

- Hadoop document and installation
  - Installation
    - http://hadoop.apache.org/core/docs/r0.20.0/quickstart.html
    - http://hadoop.apache.org/core/docs/r0.20.0/cluster_setup.html
  - Document
    - http://hadoop.apache.org/
- Hadoop Wiki
  - http://wiki.apache.org/hadoop/
- Google File System Paper
  - http://labs.google.com/papers/gfs.html

Copyright 2009 - Trend Micro Inc.