Hadoop Architecture

Philippe Julio – Principal Field Technologist
Architecture Ambassador

Sun Microsystems France

Data Management
Vision

« The Data are not created relevant, they become so ! »

Data-Driven on-Line Websites

- To run the apps : messages, posts, blog entries, video clips, maps, web graph...
- To give the data context : friends networks, social networks, collaborative filtering...
- To keep the applications running : web logs, system logs, system metrics, database query logs...

New Data and Management Economics

Compute Trend

New Analytics Emerge
(MapReduce, Hadoop…)

Google
hadoop
amazon web services™
HIVE
HBASE

Architectural shift to the cloud

Greenplum ORACLE

General purpose datawarehouse

TERADATA. ORACLE

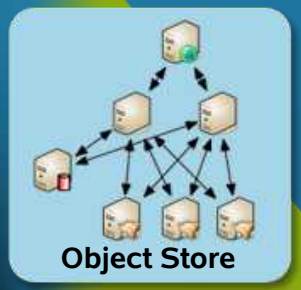
Proprietary, dedicated datawarehouse

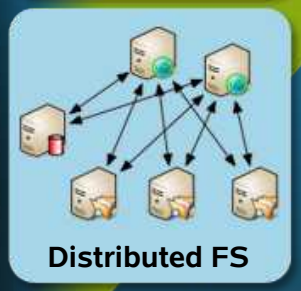
ORACLE

OLTP is the datawarehouse

Data (Storage) Trend

Semi-structured Data
(MogileFS, Bigtable, HDFS…)



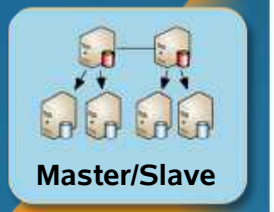


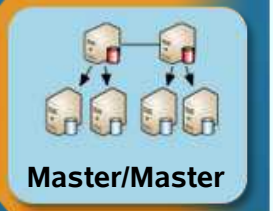
Unstructured Data

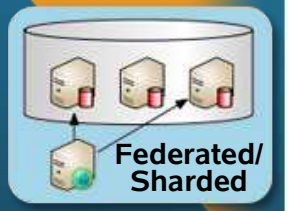
Google
hadoop

Semi-structured Database
ScaleDB, Big Table, SimpleDB hBase

MogileFS for PHP







Structured Data

What Is Hadoop ?

"Flexible and available architecture for large scale computation and data processing on a network of commodity hardware"

Open Source + Hardware Commodity = IT Costs Reduction

What Is Hadoop used for ?

- Searching
- Log processing
- Recommendation systems
- Data Warehousing
- Video and Image analysis

Who Used Hadoop ?

- Top level Apache Foundation project 
- Large, active user base, mailing lists, user groups
- Very active development, strong development team

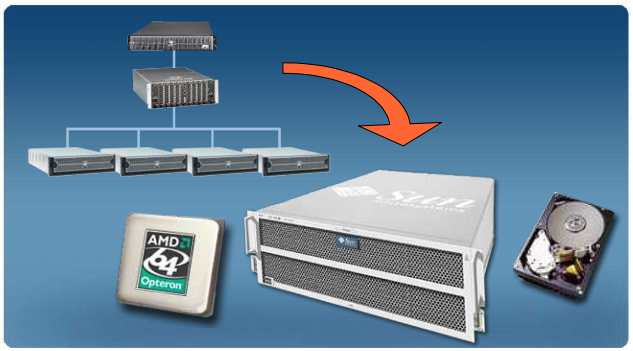
http://wiki.apache.org/hadoop/PoweredBy

Who Support Hadoop ?

- **101tec** Inc. Integration, customization, consulting. (Hadoop, Pig, Zookeeper, Lucene, Nutch)
- **Cloudera**, Inc. Get Cloudera's Distribution for Hadoop - it's free, and help you to optimize your configuration. We also provide commercial support and professional training for Hadoop. Basic training is online for free
- **Cloudify** - assist organizations in integrating Cloud Computing into their IT and Business strategies and in building and managing scalable, next-generation infrastructure environments (Hadoop, Solr, AWS, distributed architectures)
- **Doculibre** Inc. Open source and information management consulting. (Lucene, Nutch, Hadoop, Solr, Lius etc.)
- **ScaleUnlimited**, Inc. Training and mentoring on large architectures. Hadoop Bootcamp now available
- **Tinvention** -Ingegneria Informatica - Italian Consulting Company, offer support on open source architecture based on Java, including architectures based on Hadoop.

http://wiki.apache.org/hadoop/Support

Infrastructure as a Services



General Purpose Storage Servers

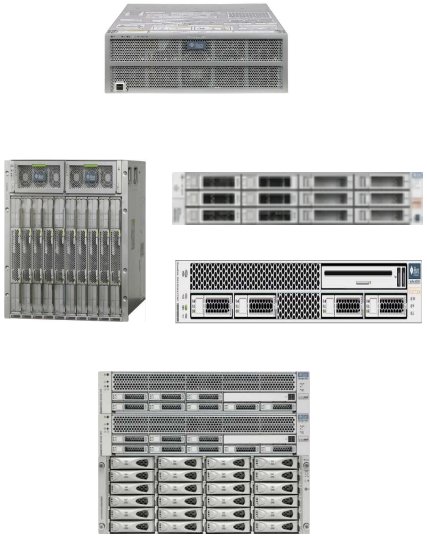
- Combine server with disk & networking
- Specialized software enables general purpose systems designs to provide high performance data services

Sun's Open Platform direction

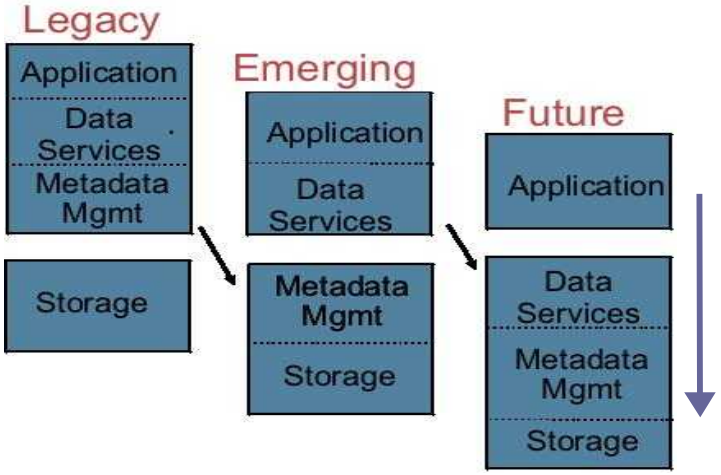
Sun Fire x4xxx
(Data Compute and Store)

Sun Sparc Enterprise T5xxx
(Data Compute and Store)

Sun Storage 7xxx
(Data Store)

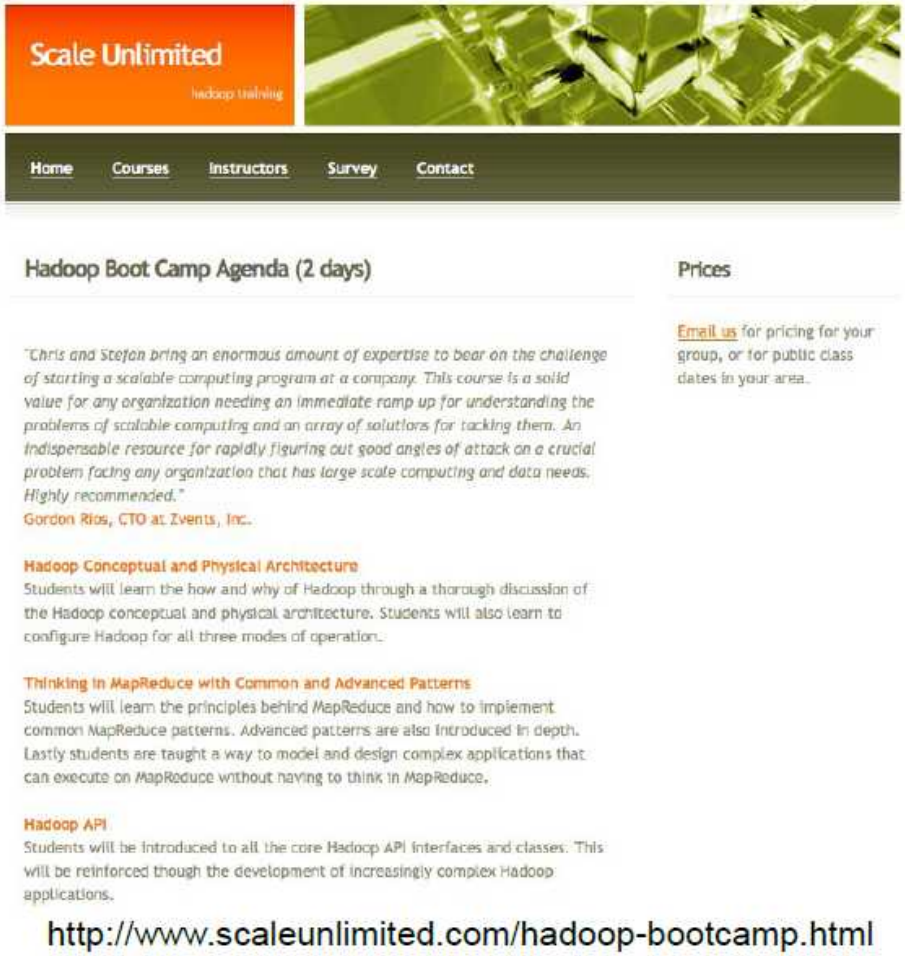


Data moves to the infrastructure



OpenSolaris Live Hadoop

- Avalaible on your Laptop
- OpenSolaris
- Hadoop Live CD based on 3 zones
  - > Global Zone : NameNode, JobTracker
  - > #1 Local Zone : DataNode, TaskTracker
  - > #2 Local Zone : DataNode, TaskTracker
- Hbase
- JDK with Java compiler and related tools
- http://opensolaris.org/os/project/livehadoop



http://www.scaleunlimited.com/hadoop-bootcamp.html

Hadoop Ecosystem

PIG (Data Flow)

HIVE (Batch SQL)

SQOOP (Data import)

ZOOKEEPER (Coordination)

AVRO (Serialization)

CHUKWA
(Displaying, Monitoring, Analyzing Logs)

MAP REDUCE (Job scheduling - Raw processing)

HBASE (Real Time Query)

HDFS
(Hadoop Distributed File System – Unstructured Storage)

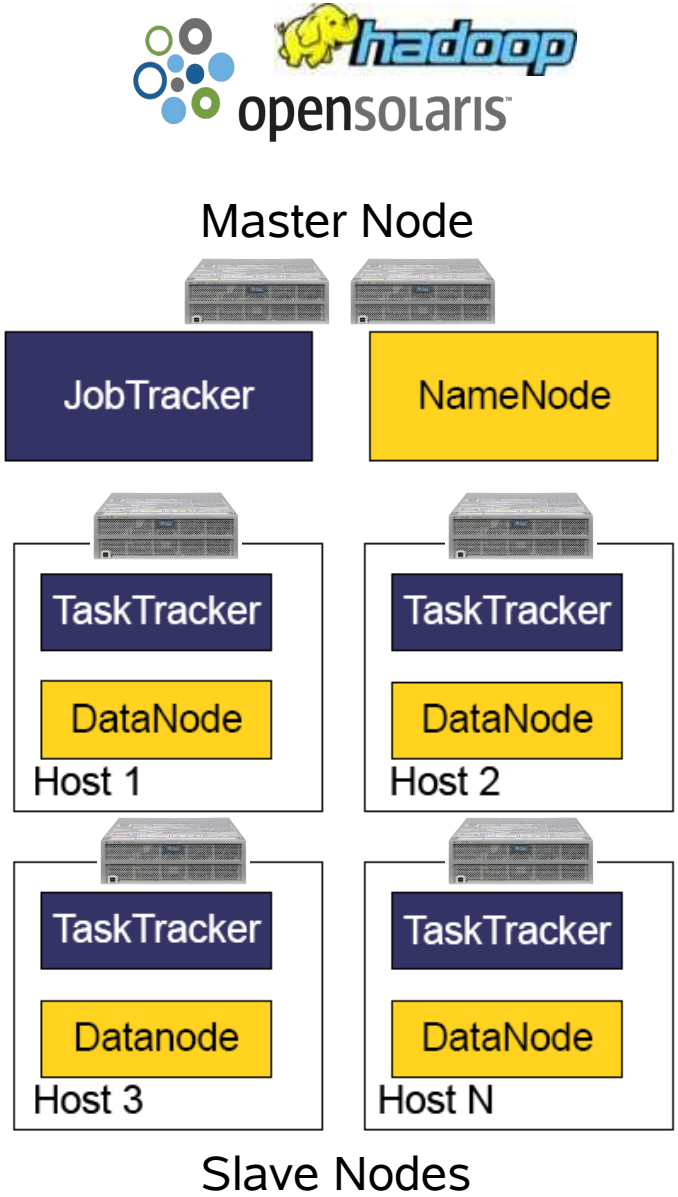
Hadoop Architecture
Hadoop Common

- Hadoop Common is a set of utilities that support the Hadoop subprojects.
- Hadoop Common includes FileSystem, RPC, and serialization libraries.

http://hadoop.apache.org/common/

Hadoop Architecture
HDFS & Map Reduce

- Hadoop Distributed File System
  - A scalable, Fault tolerant, High performance distributed file system capable of running on Sun hardware
  - Hadoop cluster with 3 nodes minimum
  - Data divided into 64MB or 128MB blocks, each block replicated 3 times (default)
  - No 15k RPM disks or RAID required
  - **NameNode** holds filesystem metadata
  - Files are broken up and spread over the **DataNodes**
- Hadoop Map Reduce
  - Software framework for distributed computation
  - Input | Map() | Copy/Sort | Reduce() | Output
  - **JobTracker** schedules and manages jobs
  - **TaskTracker** executes individual map() and reduce() tasks on each cluster node



Hadoop Architecture
HDFS

NameNode

- **Manages file system NameSpace**
  - Maps a file name to set of blocks
  - Maps a block to the DataNodes where it resides
- **Cluster configuration management**
- **Replication engine for blocks**
- **Metadata management**
  - Metadata are in main memory
  - List of files, list of blocks in each file
  - List of DataNode in each block
  - File attributes, replication factor...
- **Transaction Log**
  - Records for file creation, file deletion...

DataNode

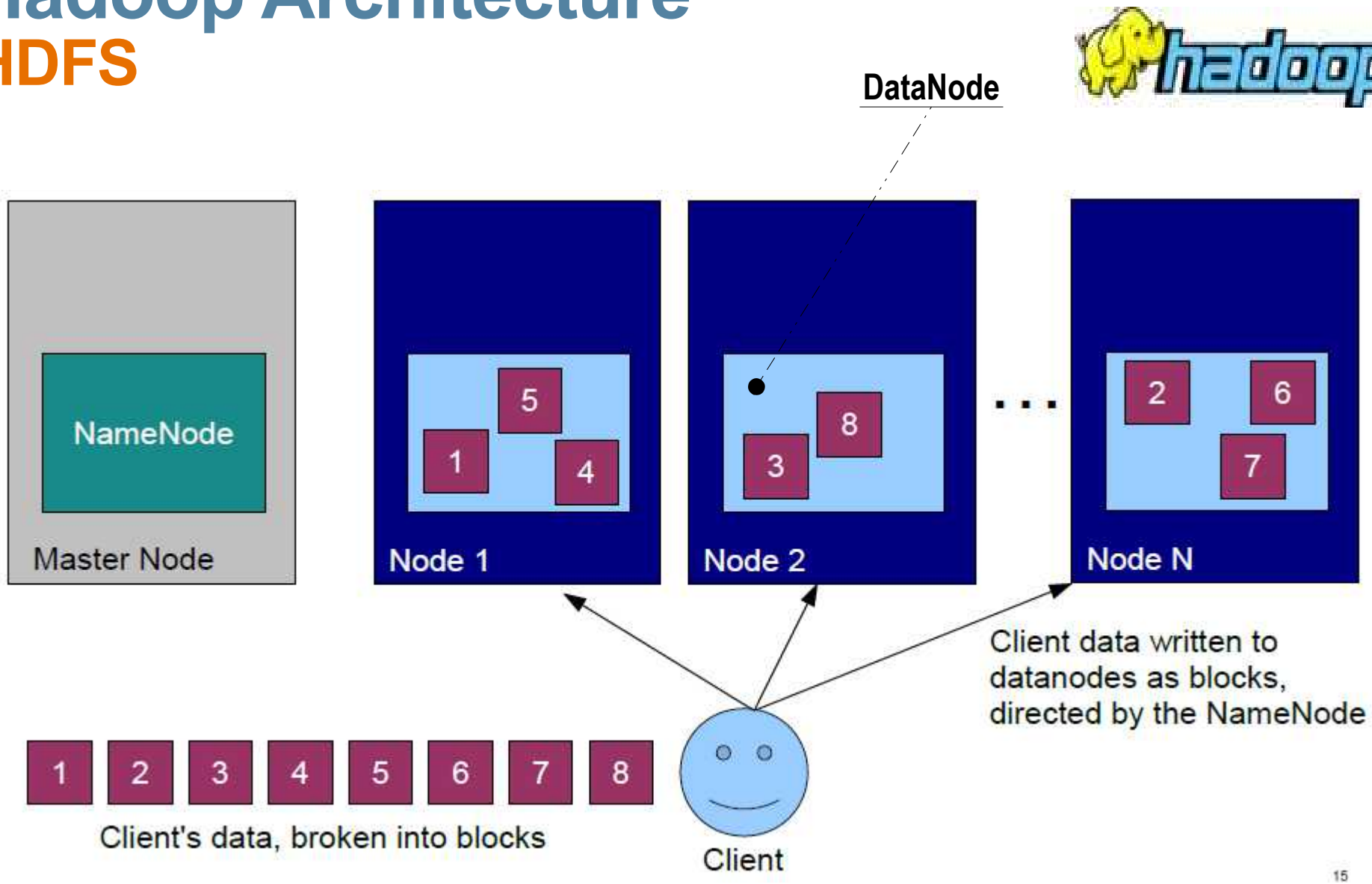
- **Block Server**
  - Stores data in the local file system
  - Stores the metadata of a block
  - Serves data and metadata to clients
- **Block Report**
  - Periodically sends a report of all existing blocks to the NameNode
- **Pipeline of Data**
  - Forwards data to other specified DataNodes

Hadoop Architecture
HDFS

- **Blocks Placement**
  - First replica on a node in a local rack
  - Second replica on different rack
  - 3rd replica on other rack
  - Clients read from nearest replica
- **Heartbeats**
  - DataNodes send heartbeat to the NameNode (once every 3 seconds)
  - NameNode used heartbeats to detect DataNode failure
- **Replication Engine**
  - Chooses new DataNodes for new replicas
  - Balances disk usage
  - Balances communication traffic to DataNodes

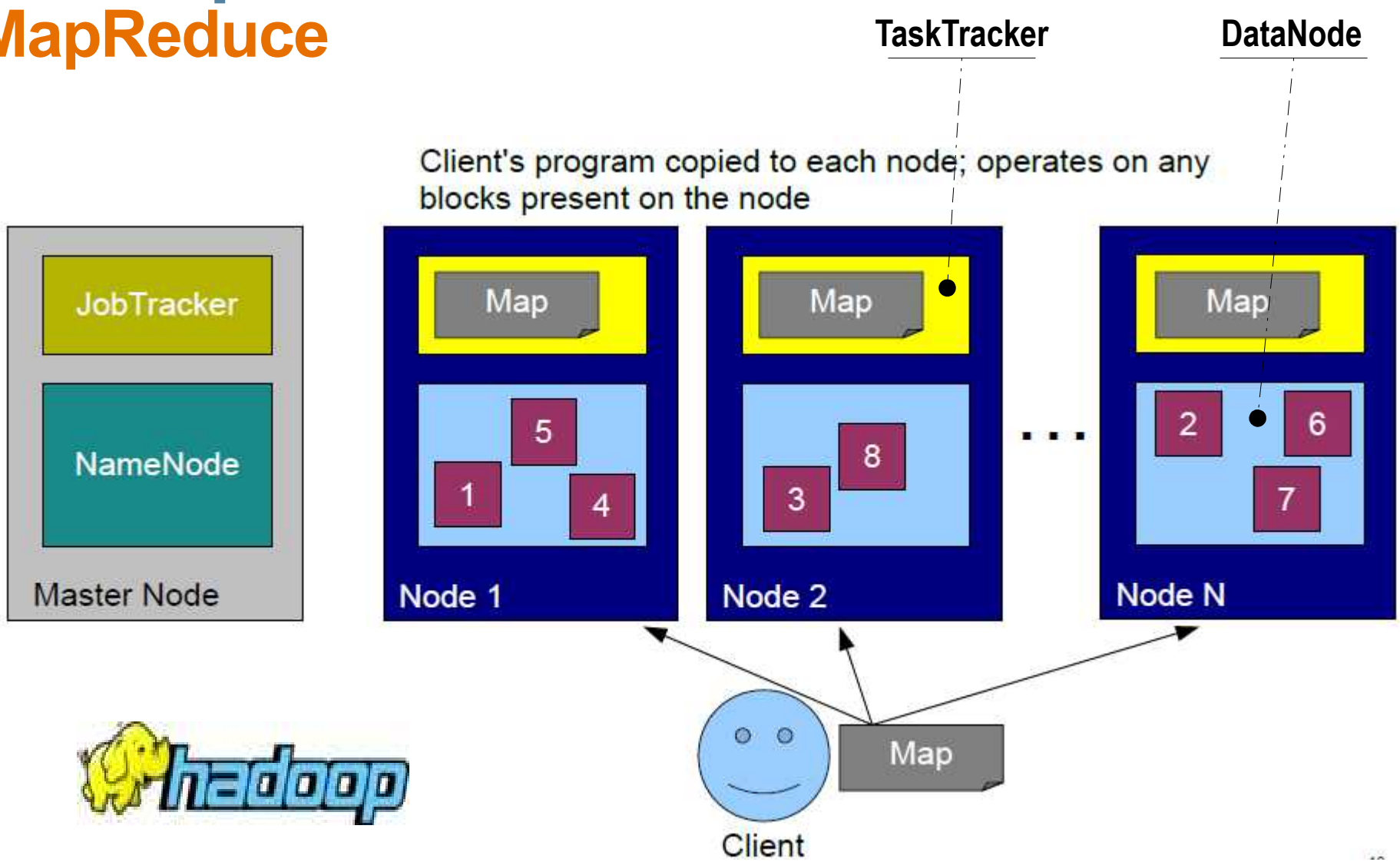
- **Data Correctness**
  - File creation : Client computes checksum per 512 bytes – DataNode stores the checksum
  - File Access : Client retrieves the data and checksum from DataNode –
    If Validation fails, Client tries other replicas
- **Data Pipeline**
  - Client retrieves a list of DataNodes on which to place replicas of a block
  - Client writes block to the first DataNode
  - The first DataNode forwards the data to the next DataNode in the Pipeline
  - When all replicas are written, the client moves on to write the next block in file
- **Rebalancer**
  - Usually run when new DataNodes are added
  - Cluster is online when Rebalancer is active
  - Rebalancer is throttled to avoid network congestion
  - Command line tool

Hadoop Architecture
HDFS



http://hadoop.apache.org/hdfs

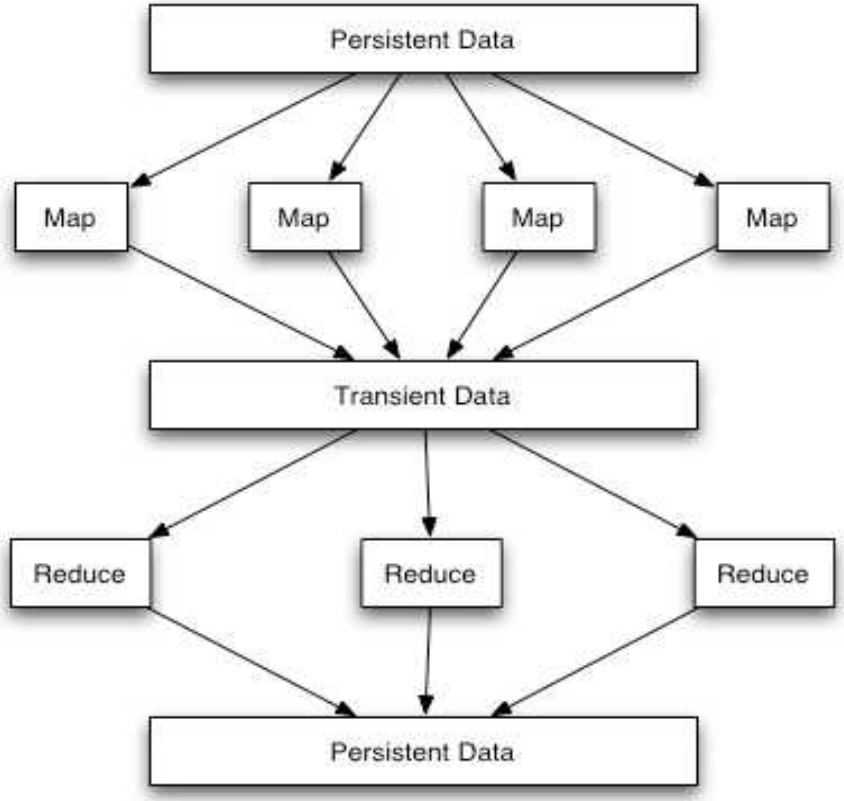
Hadoop Architecture
MapReduce



http://hadoop.apache.org/mapreduce

Hadoop Architecture
MapReduce

Input | Map() | Copy/Sort | Reduce() | Output



Map Phase

- Raw data analyzed and converted to name/value pair

Shuffle Phase

- All name/value pairs are sorted and grouped by their keys

Reduce Phase

- All values associated with a key are processed for results

Hadoop Architecture

HBase

- Clone of Big Table (Google)
- Implemented in Java (Clients : Java, C++, Ruby...)
- Data is stored "Column-oriented"
- Distributed over many servers
- Tolerant of machine failure
- Layered over HDFS
- Strong consistency

- It's not a relational database (No Joins)
- Sparse data – nulls are stored for free
- Semi-structured or unstructured data
- Data changes through time
- Versioned data
- Scalable – Goal of billions of rows x millions of columns

Table

Row	Timestamp	Animal		Repair
		Type	Size	Cost
Enclosure1	t2	Zebra	Medium	1000€
	t1	Lion	Big	
Enclosure2	t3	Monkey	Small	1500€

Region

Key

Column

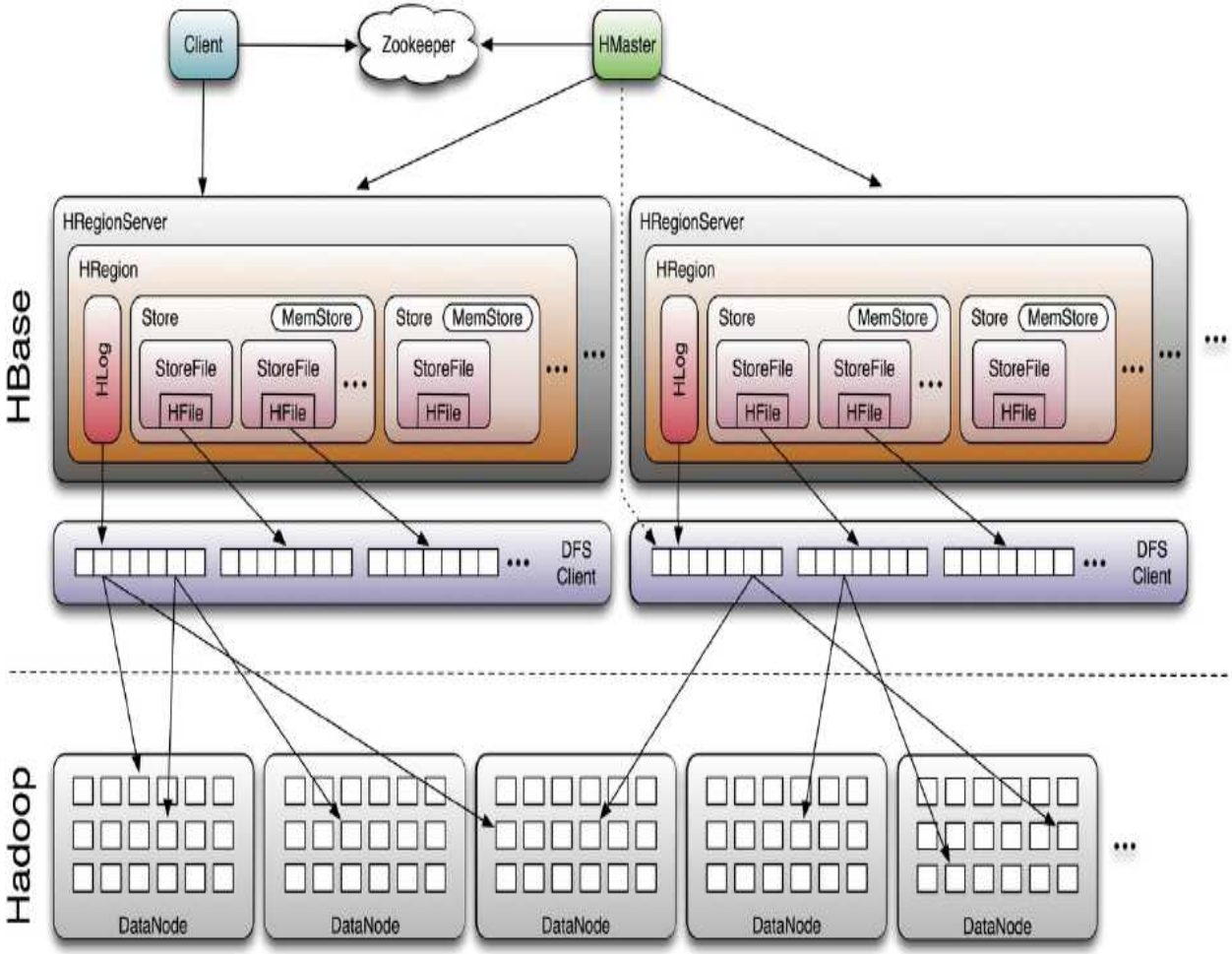
Family

Cell

(Table, Row_Key, Family, Column, Timestamp) = Cell (Value)

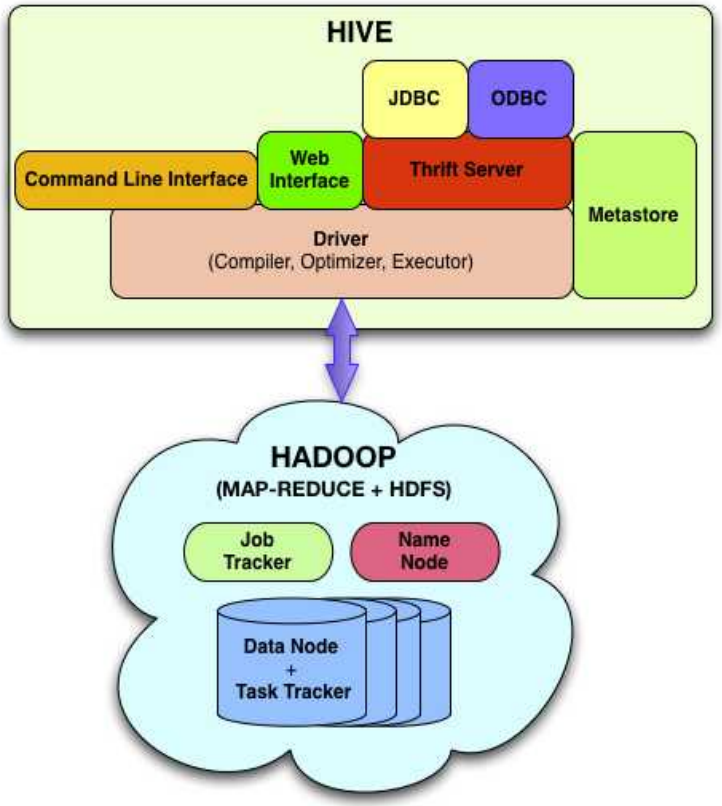
Hadoop Architecture
HBase

- Table
  - Regions for scalability, defined by row [start_key, end_key)
  - Store for efficiency, 1 per Family
    - 1..n StoreFiles (HFile format on HDFS)
- Everything is byte
- Rows are ordered sequentially by key
- Special tables -ROOT- , .META.
  - Tell clients where to find user data



http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html

Hadoop Architecture
Hive



http://hadoop.apache.org/hive

- Data Warehouse infrastructure that provides data summarization and ad hoc querying on top of Hadoop
  - MapReduce for execution
  - HDFS for storage
- MetaStore
  - Table/Partitions properties
  - Thrift API : Current clients in Php (Web Interface), Python interface to Hive, Java (Query Engine and CLI)
  - Metadata stored in any SQL backend
- Hive Query Language
  - Basic SQL : Select, From, Join, Group By
  - Equi-Join, Multi-Table Insert, Multi-Group-By
  - Batch query

Hadoop Architecture
Pig

- A high-level data-flow language and execution framework for parallel computation
- Simple to write MapReduce program
- Abstacts you from specific detail
- Focus on data processing
- Data flow
- For data manipulation

http://hadoop.apache.org/pig

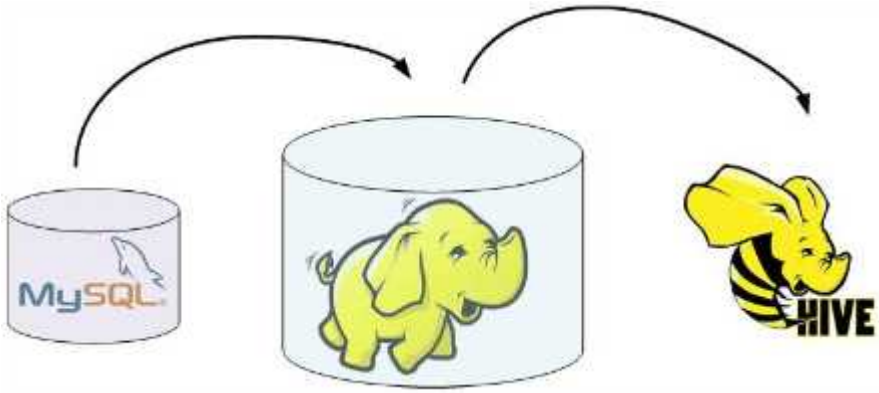
PIG Language Example

```
Users = LOAD 'users' AS (name, age);
Fltrd = FILTER Users BY
        age >= 18 AND age <= 25;
Pages = LOAD 'pages' AS (user, url);
Jnd = JOIN Fltrd BY name, Pages BY user;
Grpd = GROUP Jnd BY url;
Smmd = FOREACH Grpd GENERATE group,
       COUNT(Jnd) AS clicks;
Srtd = ORDER Smmd BY clicks DESC;
Top5 = LIMIT Srtd 5;
STORE Top5 INTO 'top5sites';
```

Hadoop Architecture
Sqoop

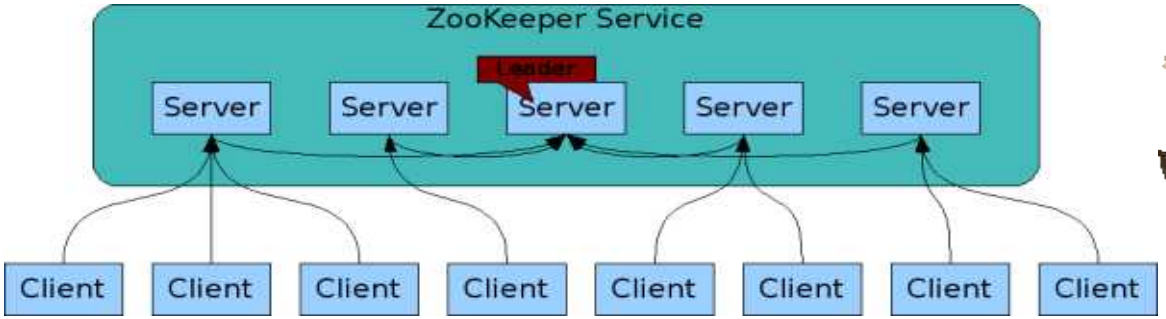
- Sqoop is a tool designed to help users of large data import existing relational databases into their Hadoop clusters
- Automatic data import
- SQL-to-Hadoop
- Easy import data from many databases to Hadoop
- Generates code for use in MapReduce applications
- Integrates with Hive

http://www.cloudera.com/hadoop-sqoop



Hadoop Architecture
Zookeeper

- A high-performance coordination service for distributed applications
- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services



- All servers store a copy of the data
- A leader is elected at startup
- Followers service clients, all updates go through leader
- Update responses are sent when a majority of servers have persisted the change

http://hadoop.apache.org/zookeeper

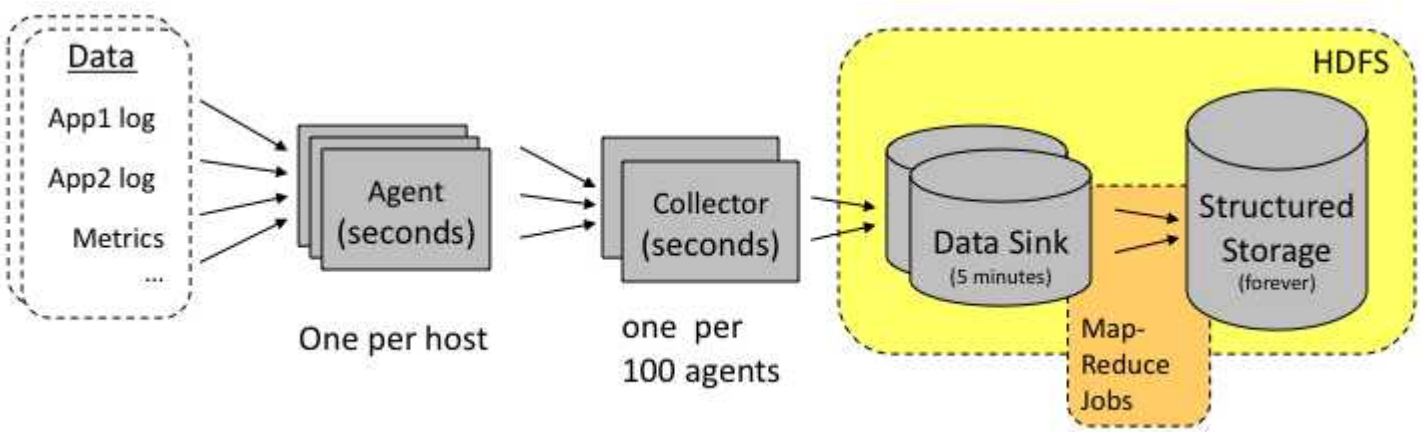
Hadoop Architecture
Avro

- A data serialization system that provides dynamic integration with scripting languages
- Avro Data
  - Expressive
  - Smaller and Faster
  - Dynamic
    - Schema store with data
    - APIs permit reading and creating
  - Include a file format and a textual encoding
- Avro RPC
  - Leverage versioning support
  - For Hadoop service provide cross-language access

http://hadoop.apache.org/avro/docs/current

Hadoop Architecture
Chukwa

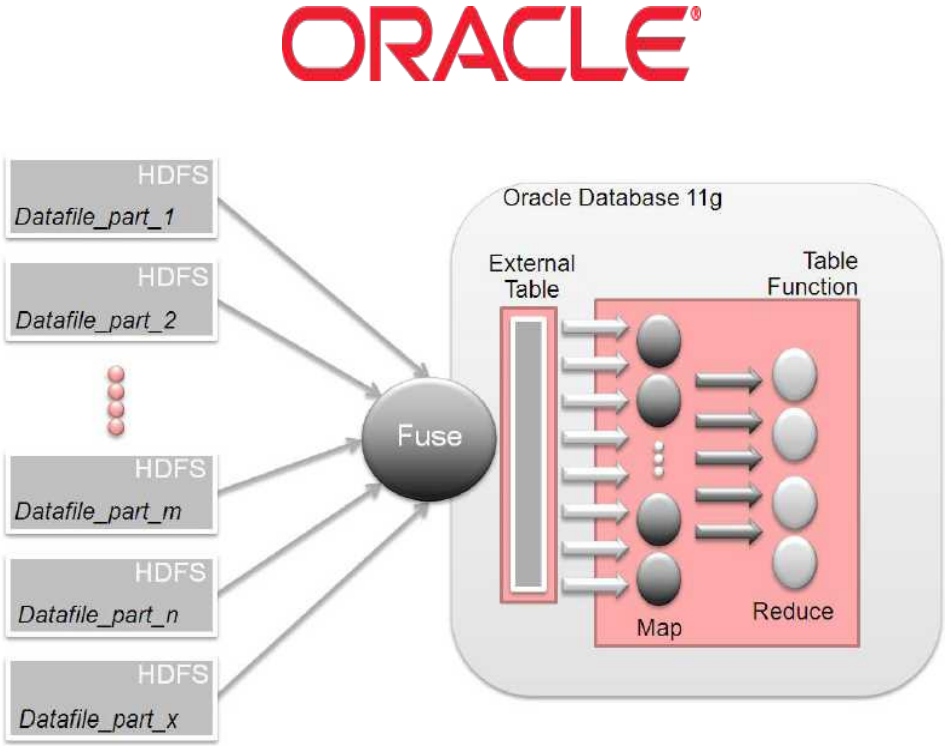
- A data collection system for managing large distributed systems
- Build on HDFS and MapReduce
- Tools kit for displaying, monitoring and analyzing the log files



http://hadoop.apache.org/chukwa

Hadoop Architecture
Oracle & HDFS

- External tables present data stored in a file system in a table format
- Use SQL queries transparently
- The FUSE : **F**ile system in **USE**rspace
- FUSE drivers allow users to mount a HDFS store and treat it like a normal file system
- Oracle Table Functions provide an alternate way to fetch data from Hadoop



Hadoop Architecture
x86 Servers Components

Low Cost Server & Storage : Sun Fire X4xxx

Sun Virtualization Technology : Solaris Containers

Sun Fire X4140
- 2 CPU Quad Core AMD
- Up to 128GB RAM
- Up to 8 Disks
- Up to 2,3TB Disks
- 1 RU

Sun Fire X4240
- 2 CPU Quad Core AMD
- Up to 128GB RAM
- Up to 16 Disks
- Up to 4,6TB Disks
- 2 RU

Sun Fire X4275
- 2 CPU Quad Core Intel
- Up to 144GB RAM
- Up to 12 Disks
- Up to 24TB Disks
- 2 RU

Sun Fire X4540
- 2 CPU Quad CoreAMD
- Up to 64GB RAM
- Up to 48 Disks
- Up to 96TB Disks
- 4 RU

Interface
- HDFS
- NFS
- HTTP
- ...

Hadoop Architecture
x86 Servers Components

Low Cost Server : Sun Blade

Sun Virtualization Technologies : Solaris Containers

Sun Blade 6000
- Up to 10 Blades
- 10 RU

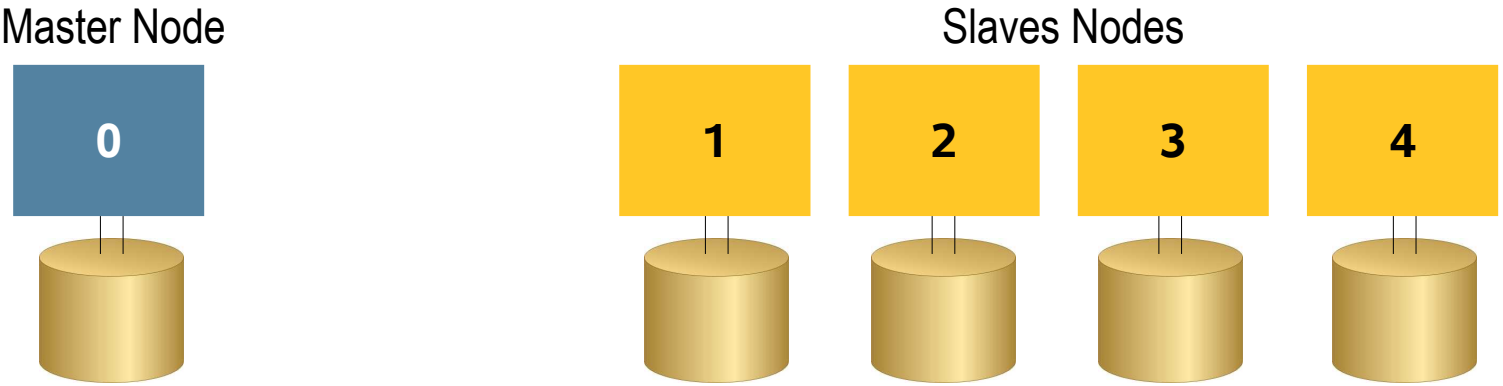
Sun Blade 6270
- 2 CPU Quad Core Intel
- Up to 144GB RAM
- Up to 4 Disks
- Up to 2TB Disks

Interface
- HDFS
- NFS
- HTTP
- ...

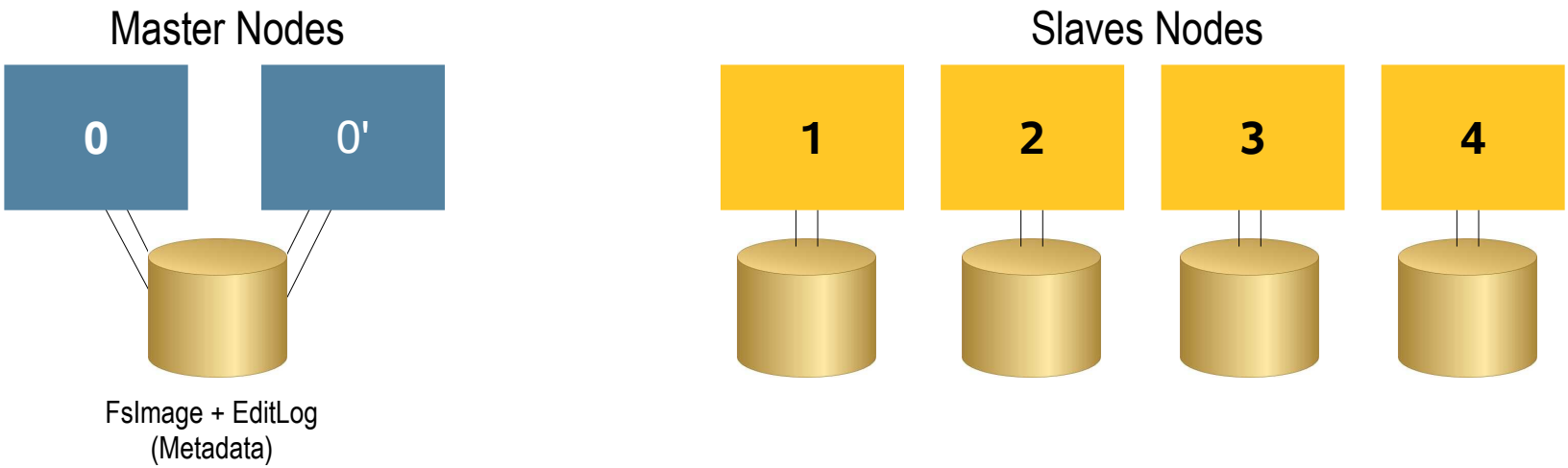
Hadoop Architecture
High Availability

Hadoop Cluster

Single Point Of Failure (SPOF)

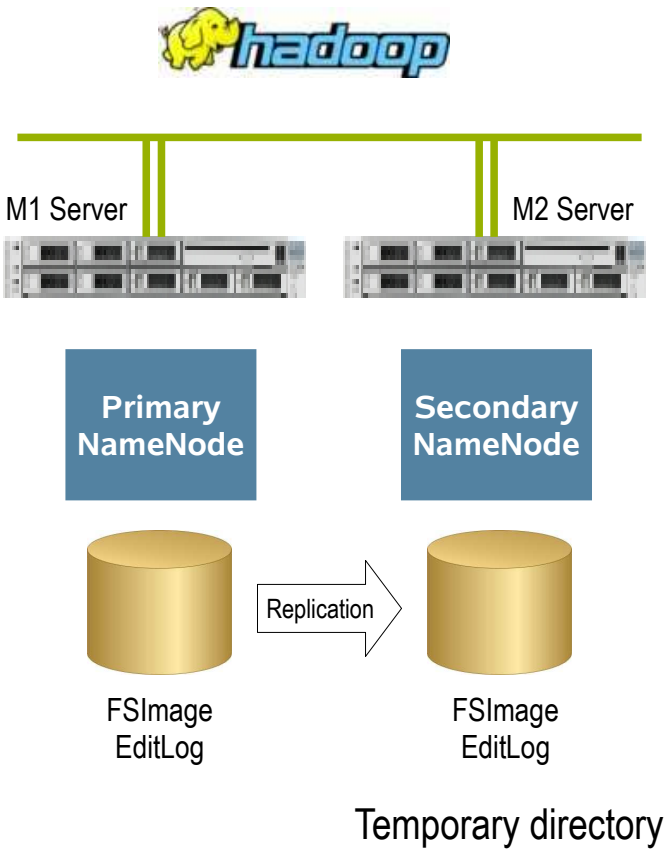


High Availability Hadoop Cluster



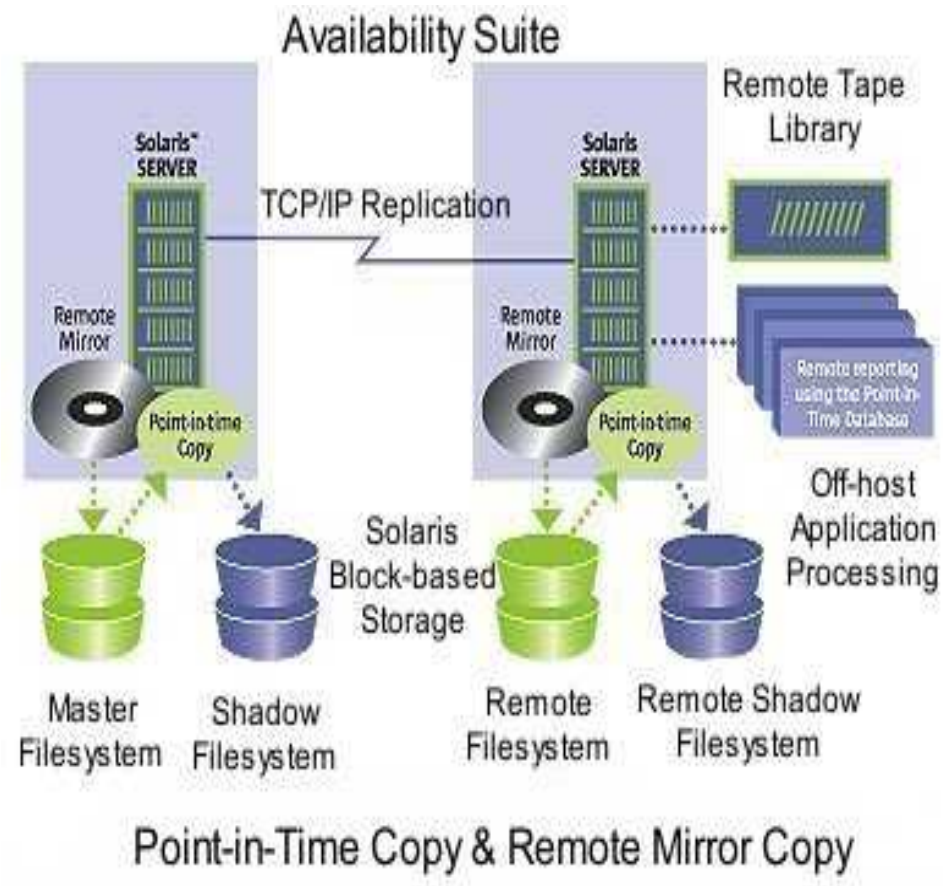
Hadoop Architecture
High Availability with Secondary NameNode

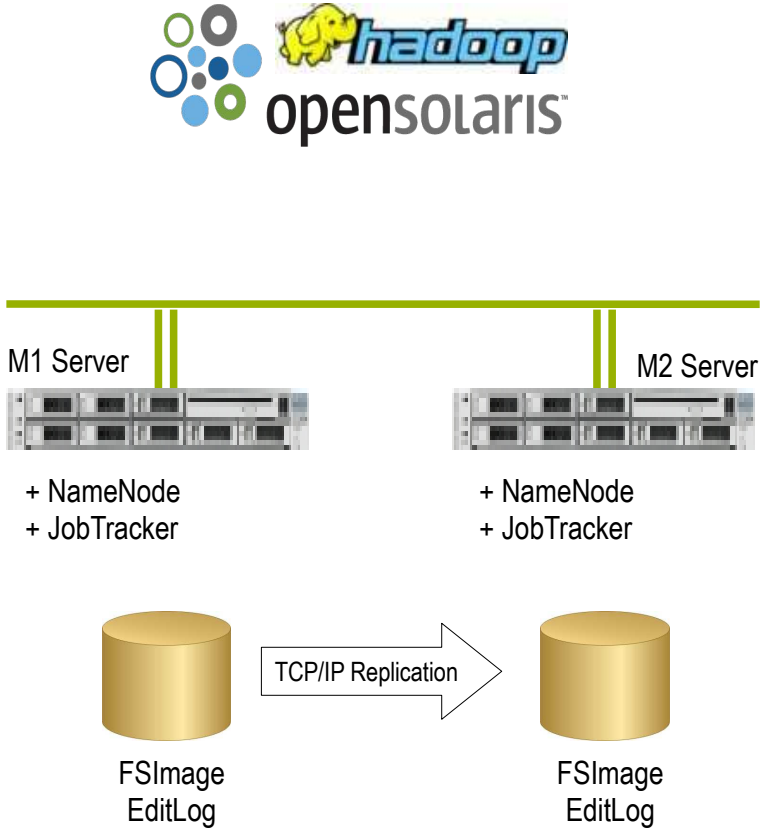
- Is usually run on different servers
  - Primary et Secondary NameNode
- Copies FSImage and transaction Log (EditLog) from NameNode to a temporary directory
- Merges FSImage and Transaction Log periodically into a new FSImage in temporary directory
- Uploads new FSImage to the NameNode
- Purges transaction Log on the NameNode



Hadoop Architecture

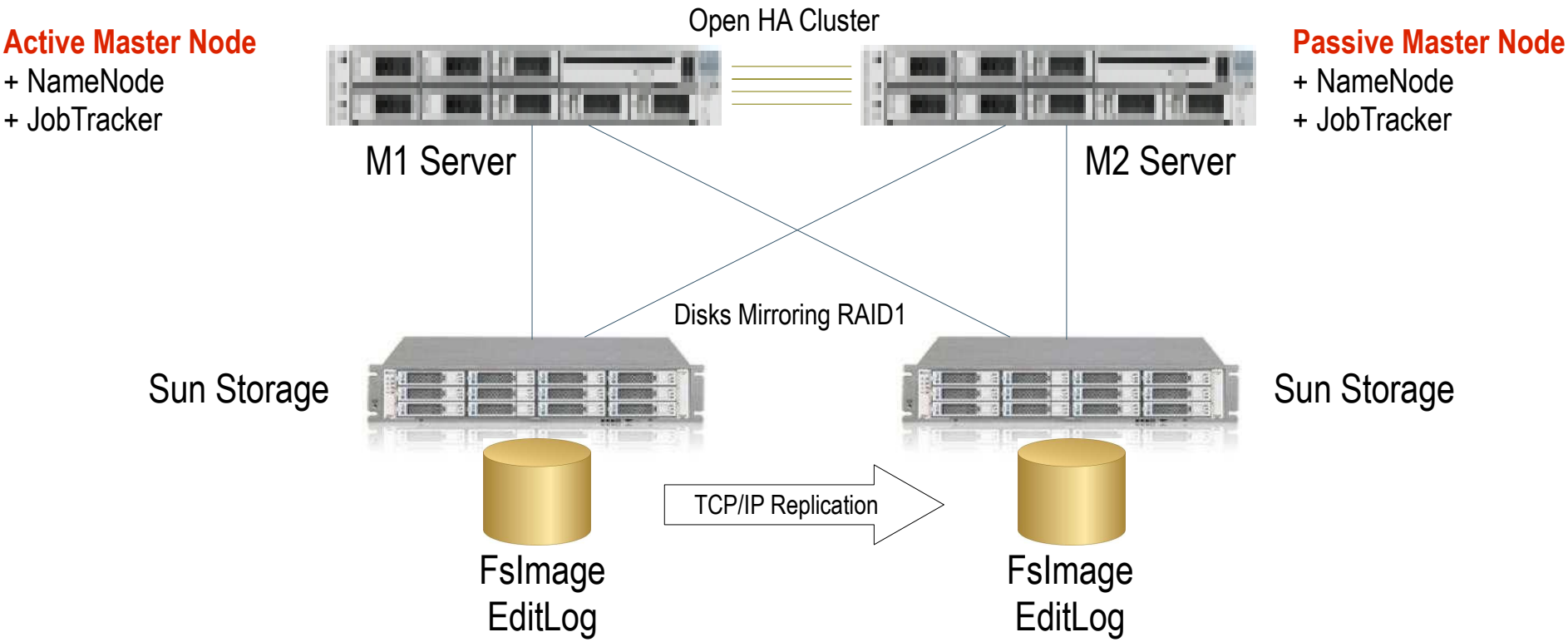
High Availability with Sun StorageTek Availability Suite



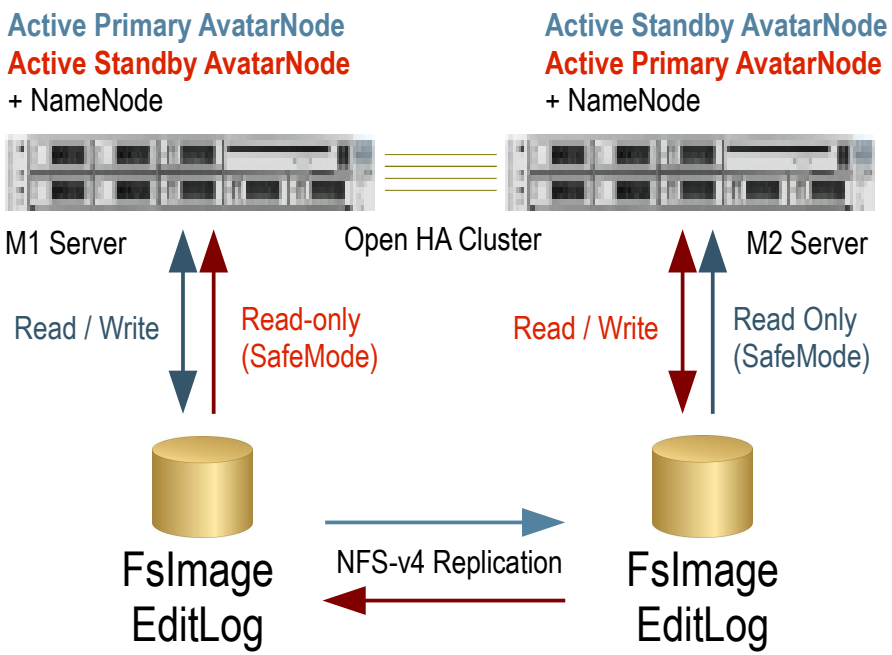


http://hub.opensolaris.org/bin/view/Project+avs/WebHome

Hadoop Architecture
High Availability with Open HA Cluster



Hadoop Architecture
High Availability with AvatarNode

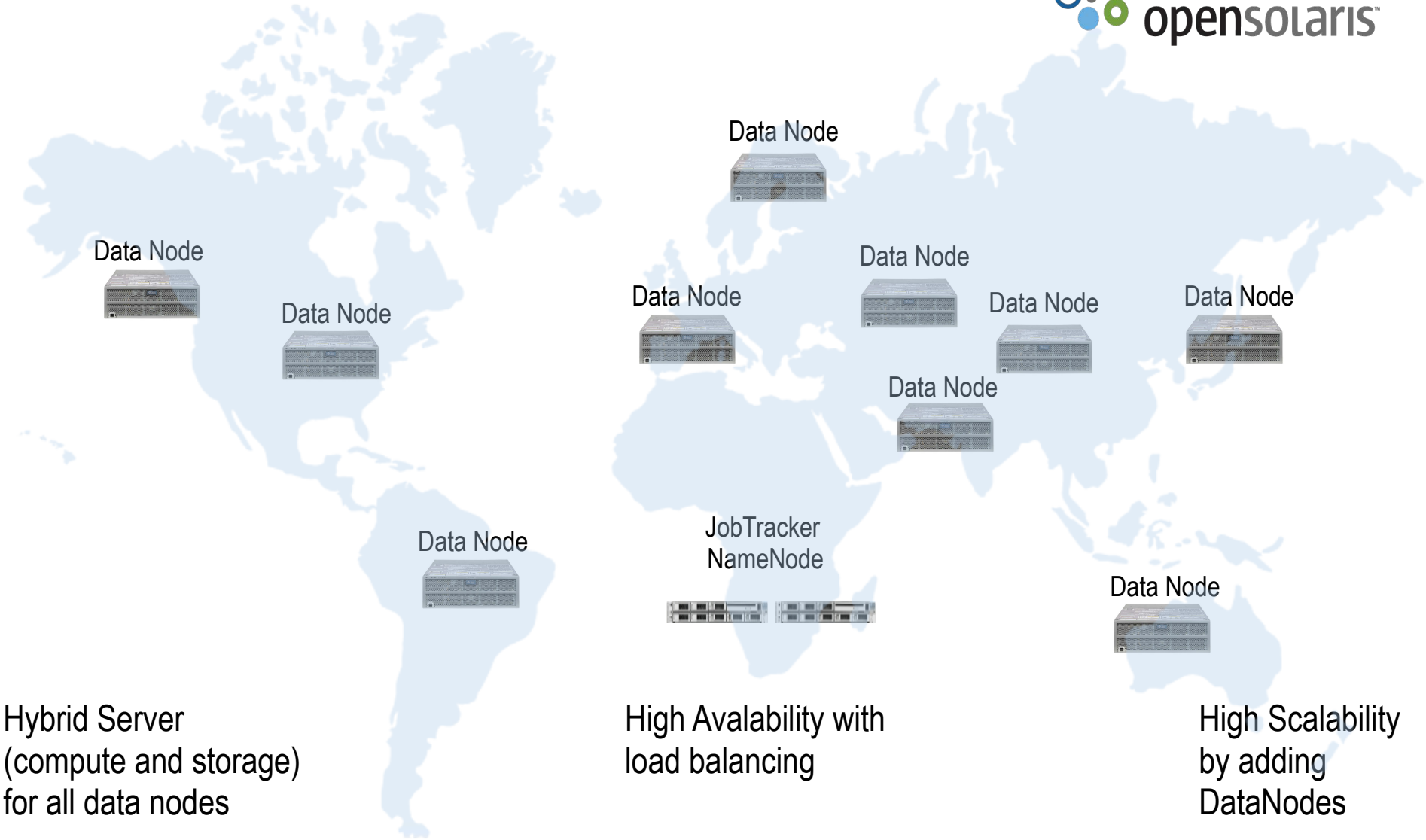


- HDFS clients are configured to access the AvatarNode via a Virtual IP Address (VIP)
- When PrimaryAvatarNode is down, the Standby AvatarNode takes the relay
- The Standby AvatarNode ingests all committed transactions because it reopens the edits log and consumes all transactions until the end of the file
- The Standby AvatarNode finishes ingestion of all transactions from the shared NFS filer and then leaves SafeMode
- The VIP switches from Primary AvatarNode to Standby AvatarNode

http://hadoopblog.blogspot.com/2010/02/hadoop-namenode-high-availability.html

- Code has been contributed to the Apache HDFS project via HDFS-976. A prerequisite for this patch is HDFS-966.

Hadoop Architecture
World Wide IP Cluster HA



Hybrid Server
(compute and storage)
for all data nodes

High Avalability with
load balancing

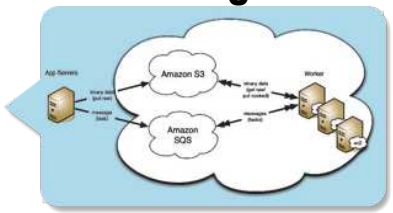
High Scalability
by adding
DataNodes

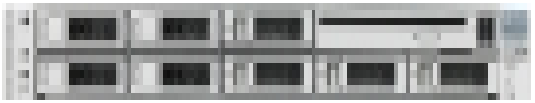
Hadoop Architecture
Sizing

Assumptions

- Business Data Volume = Customer needs
- No RAID factor, No HBA port
- 2 CPU Quad-core for all servers
- 2 System hard disks
- Number of replication blocks = 3
- Block size = 128 MB

Cloud Storage Model



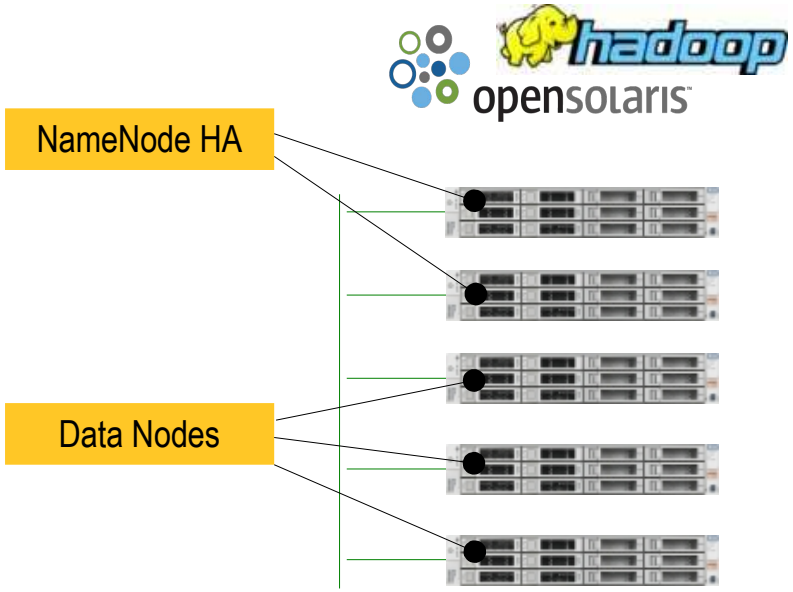


Sizing for HA Cluster

- Temporary Space = 25% of the total hard disk
- Raw Data Volume = 1.25 * (Business Data Volume * Nb of replication blocks)
- Number of NameNode Servers = 2
- Number of DataNode Servers = Raw Data Volume / Server Capacity Storage
- NameNode RAM = 64 GB
- DataNode RAM = 32 GB mini

Hadoop Architecture
Proof Of Concept

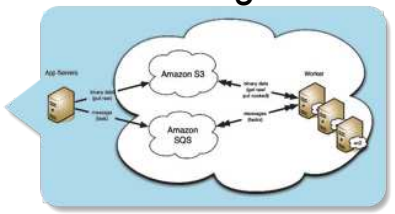
- 1 Primary NameNode
- 1 Secondary NameNode
- 3 Data Nodes



Sun Fire X4275

- 2 CPU Quad Core Intel
- Up to 144GB RAM
- 2 Boot Disks
- 10 Disks for Hadoop
- Up to 24TB Disks
- 2 RU

Cloud Storage Model



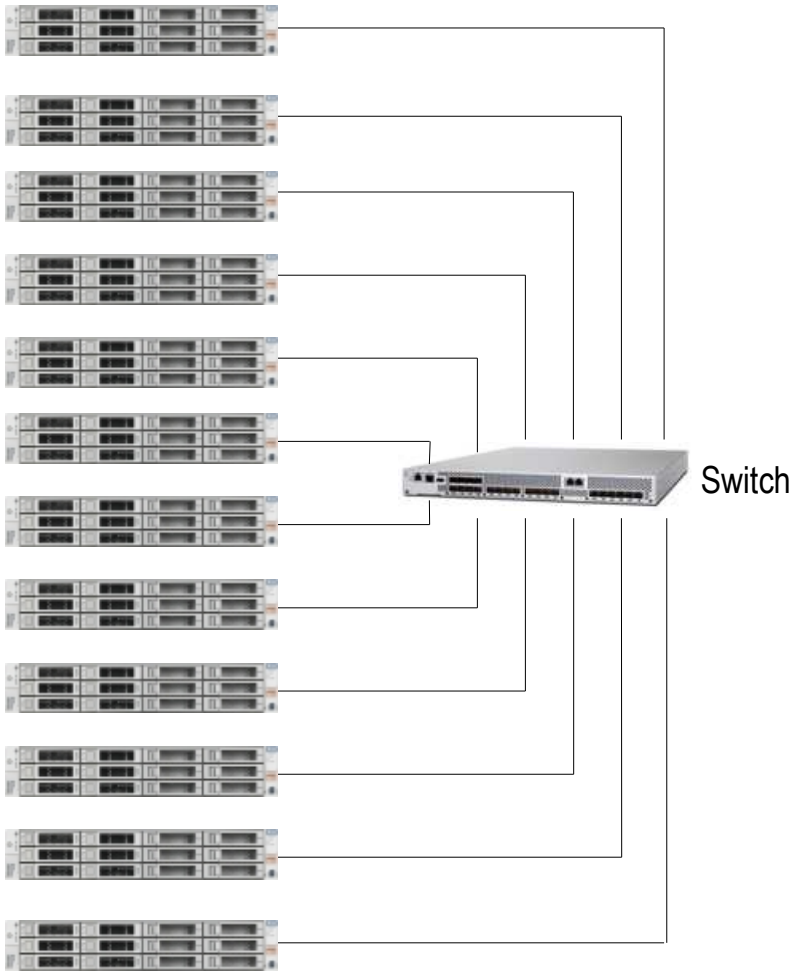
Hadoop Architecture
Production

Master Nodes

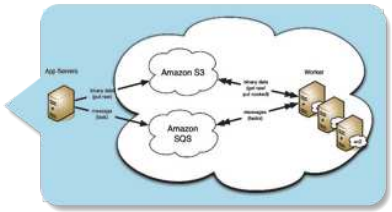
Sun Fire X4275

- 2 CPU Quad Core Intel
- Up to 144GB RAM
- 2 Boot Disks
- 10 Disks for Hadoop
- Up to 24TB Disks
- 2 RU

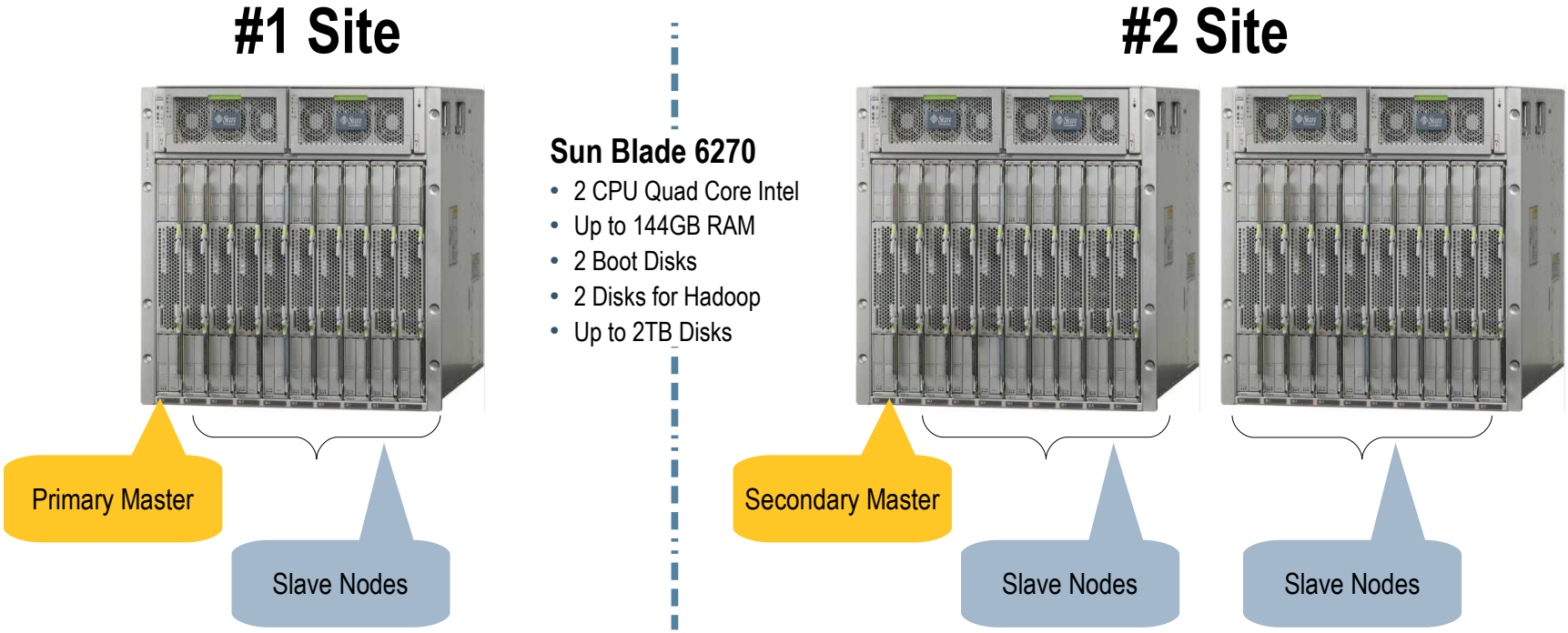
Slave Nodes



Cloud Storage Model

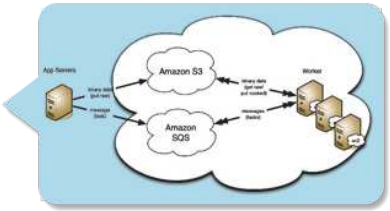


Hadoop Architecture
Production



- 1 Primary NameNode / Rack
- 1 Secondary NameNode / Rack
- 9 DataNode / Rack
- 1 Blade 6270 / Node
- For other racks : 10 DataNode / Rack

Cloud Storage Model



Philippe Julio
philippe.julio@sun.com
http://blogs.sun.com/philippejulio