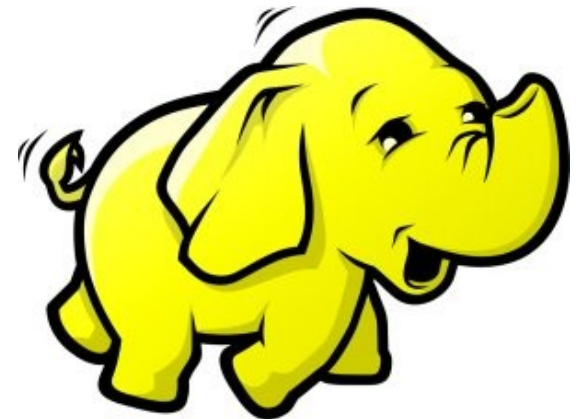




# Hadoop 相關計畫

*Hadoop Ecosystem*

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**

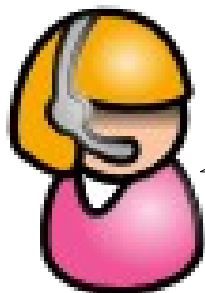




**Hadoop** 只支援用 **Java** 開發嘛？  
***Is Hadoop only support Java ?***

總不能全部都重新設計吧？如何與舊系統相容？

***Can Hadoop work with existing software ?***



可以跟資料庫結合嘛？

***Can Hadoop work with Databases ?***

開發者們有聽到大家的需求.....

***Yes, we hear the feedback of developers ...***



# Is Hadoop only support Java ?

- Although the Hadoop framework is implemented in Java<sup>™</sup>, **Map/Reduce applications need not be written in Java.**
- **Hadoop Streaming** is a utility which allows users to **create and run jobs with any executables (e.g. shell utilities)** as the mapper and/or the reducer.
- **Hadoop Pipes** is a SWIG-compatible **C++ API** to implement Map/Reduce applications (non JNI<sup>™</sup> based).

# Hadoop Pipes (C++, Python)

- Hadoop Pipes allows **C++** code to use Hadoop DFS and map/reduce.
- The C++ interface is "swigable" so that interfaces can be generated for **python** and other scripting languages.
- For more detail, check the API Document of [org.apache.hadoop.mapred.pipes](http://org.apache.hadoop.mapred.pipes)
- You can also find example code at [hadoop-\\*/src/examples/pipes](http://hadoop-*/src/examples/pipes)
- About the pipes C++ WordCount example code: <http://wiki.apache.org/hadoop/C++WordCount>

# Hadoop Streaming

- Hadoop Streaming is a utility which allows users to create and run Map-Reduce jobs **with any executables (e.g. Unix shell utilities)** as the mapper and/or the reducer.
- It's useful when you need to run **existing program** written in shell script, perl script or even PHP.
- Note: both the **mapper** and the **reducer** are **executables** that read the input from **STDIN** (line by line) and emit the output to **STDOUT**.
- For more detail, check the official document of **Hadoop Streaming**

# Running Hadoop Streaming

```
jazz@hadoop:~$ hadoop jar hadoop-streaming.jar -help
10/08/11 00:20:00 ERROR streaming.StreamJob: Missing required option -input
Usage: $HADOOP_HOME/bin/hadoop [--config dir] jar \
    $HADOOP_HOME/hadoop-streaming.jar [options]
```

Options:

```
-input      <path>          DFS input file(s) for the Map step
-output     <path>          DFS output directory for the Reduce step
-mapper     <cmd|JavaClassName>    The streaming command to run
-combiner   <JavaClassName> Combiner has to be a Java class
-reducer    <cmd|JavaClassName>    The streaming command to run
-file       <file>          File/dir to be shipped in the Job jar file
-dfs        <h:p>|local  Optional. Override DFS configuration
-jt         <h:p>|local  Optional. Override JobTracker configuration
-additionalconfspec specfile  Optional.
-inputformat TextInputFormat(default) |SequenceFileAsTextInputFormat |
JavaClassName Optional.
-outputformat TextOutputFormat(default) |JavaClassName  Optional.
```

... More ...

# Hadoop Streaming with shell commands (1)

```
hadoop:~$ hadoop fs -rmr input output
```

```
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
```

```
hadoop:~$ hadoop jar hadoop-streaming.jar -input  
input -output output -mapper /bin/cat -reducer /  
usr/bin/wc
```

# Hadoop Streaming with shell commands (2)

```
hadoop:~$ echo "sed -e \"s/ /\n/g\" | grep ." >  
streamingMapper.sh
```

```
hadoop:~$ echo "uniq -c | awk '{print \$2 \"\t\"  
\$1}'" > streamingReducer.sh
```

```
hadoop:~$ chmod a+x streamingMapper.sh
```

```
hadoop:~$ chmod a+x streamingReducer.sh
```

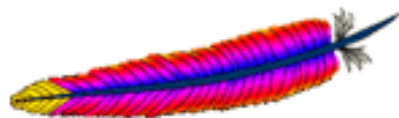
```
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
```

```
hadoop:~$ hadoop jar hadoop-streaming.jar -input  
input -output output -mapper streamingMapper.sh  
-reducer streamingReducer.sh -file  
streamingMapper.sh -file streamingReducer.sh
```



# There are several Hadoop subprojects

Apache > Hadoop >



Top

Common

Chukwa

HBase

HDFS

Hive

MapReduce

Pig

ZooKeeper

▼ About

▫ Welcome

▫ Who We Are?

▫ Mailing Lists

## Welcome to Apache Hadoop!

- **Hadoop Common:** The common utilities that support the other Hadoop subprojects.
- **HDFS:** A distributed file system that provides high throughput access to application data.
- **MapReduce:** A software framework for distributed processing of large data sets on compute clusters.

## Other Hadoop related projects

- **Chukwa**: A data collection system for managing large distributed systems.
- **HBase**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Pig**: A high-level data-flow language and execution framework for parallel computation.
- **ZooKeeper**: A high-performance coordination service for distributed applications.

# Hadoop Ecosystem

<b><i>Pig</i></b>	<b><i>Chukwa</i></b>	<b><i>Hive</i></b>	<b><i>HBase</i></b>
<b><i>MapReduce</i></b>		<b><i>HDFS</i></b>	<b><i>ZooKeeper</i></b>
<b><i>Hadoop Core (Hadoop Common)</i></b>		<b><i>Avro</i></b>	

Source: *Hadoop: The Definitive Guide*

# Avro

- Avro is a **data serialization system**.
- It provides:
  - *Rich data structures.*
  - *A compact, fast, binary data format.*
  - *A container file, to store persistent data.*
  - *Remote procedure call (RPC).*
  - *Simple integration with dynamic languages.*
- Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth implementing for statically typed languages.
- For more detail, please check the official document:  
<http://avro.apache.org/docs/current/>



# Zoo Keeper



- <http://hadoop.apache.org/zookeeper/>
- ZooKeeper is a **centralized service** for **maintaining configuration** information, **naming**, **providing distributed synchronization**, and providing group services. All of these kinds of services are used in some form or another by distributed applications.
- *Each time they are implemented there is a lot of work that goes into fixing the bugs and **race conditions** that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.*

# Pig

- <http://hadoop.apache.org/pig/>
- Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
- Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs
- Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:
  - Ease of programming
  - Optimization opportunities
  - Extensibility



# Hive

- <http://hadoop.apache.org/hive/>
- Hive is a **data warehouse** infrastructure built on top of Hadoop that provides tools to enable easy **data summarization**, **adhoc querying** and analysis of large datasets data stored in Hadoop files.
- **Hive QL** is based on SQL and enables users familiar with SQL to query this data.



# Chukwa

- <http://hadoop.apache.org/chukwa/>
- Chukwa is an open source **data collection system** for monitoring large distributed systems.
- built on top of HDFS and Map/Reduce framework
- includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.





# Mahout

- <http://mahout.apache.org/>
- Mahout is a scalable **machine learning libraries**.
- implemented on top of Apache Hadoop using the map/reduce paradigm.
- Mahout currently has
  - Collaborative Filtering
  - User and Item based recommenders
  - **K-Means, Fuzzy K-Means clustering**
  - Mean Shift clustering
  - More ...

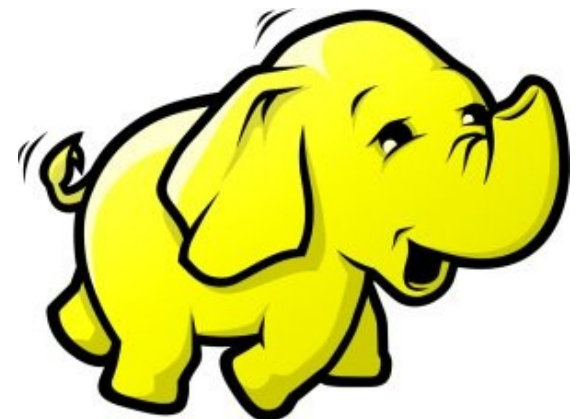




# HBase 雲端資料庫

*Introduction to HBase*

***Jazz Wang***  
***Yao-Tsung Wang***  
***jazz@nchc.org.tw***



# It's all about SCALE!!



**Warning:** fopen(/home/dodgers/public\_html/./logs/oracle\_error\_log.txt) [function.fopen]: failed to open stream: Permission denied in /usr/local/apache/htdocs/include2007/oracle/db\_oracle.inc.php on line 194

Cannot open Database Error Log, please check!! (/home/dodgers/public\_html/./logs/oracle\_error\_log.txt)

**Warning:** fopen(/home/dodgers/public\_html/./logs/oracle\_error\_log.txt) [function.fopen]: failed to open stream: Permission denied in /usr/local/apache/htdocs/include2007/oracle/db\_oracle.inc.php on line 194

Cannot open Database Error Log, please check!! (/home/dodgers/public\_html/./logs/oracle\_error\_log.txt)

**Warning:** fopen(/home/dodgers/public\_html/./logs/oracle\_error\_log.txt) [function.fopen]: failed to open stream: Permission denied in /usr/local/apache/htdocs/include2007/oracle/db\_oracle.inc.php on line 194

Cannot open Database Error Log, please check!! (/home/dodgers/public\_html/./logs/oracle\_error\_log.txt)

**Warning:** fopen(/home/dodgers/public\_html/./logs/oracle\_error\_log.txt) [function.fopen]: failed to open stream: Permission denied in /usr/local/apache/htdocs/include2007/oracle/db\_oracle.inc.php on line 194

Cannot open Database Error Log, please check!! (/home/dodgers/public\_html/./logs/oracle\_error\_log.txt)



訂購歷史紀錄

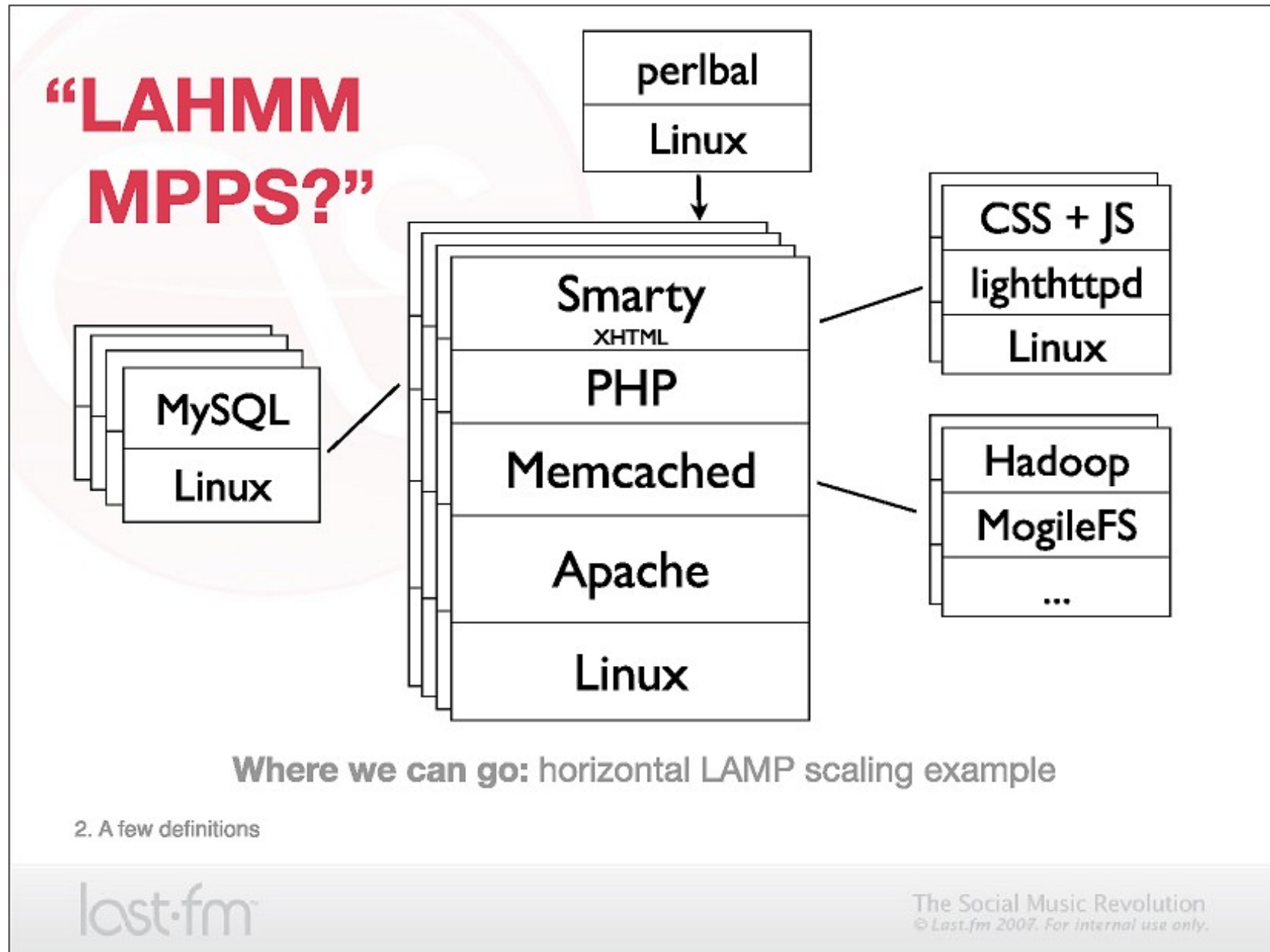


denied in /usr/local/apache/htdocs/include2007/oracle/db\_oracle.inc.php on line 194

Cannot open Database Error Log, please check!! (/home/dodgers/public\_html/./logs/oracle\_error\_log.txt)

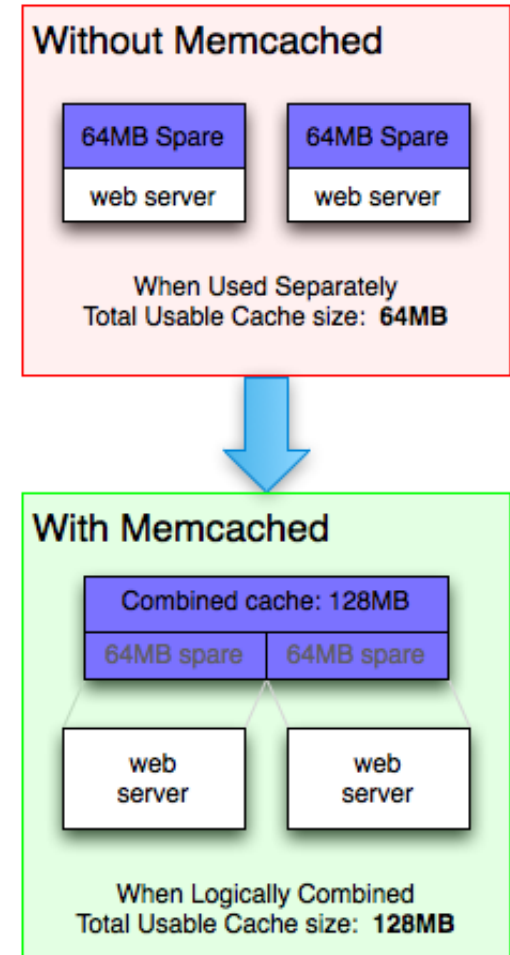
**Warning:** fopen(/home/dodgers/public\_html/./logs/oracle\_error\_log.txt) [function.fopen]: failed to open stream: Permission

# How to scale up web service in the past ?



# Tools used by large scale websites

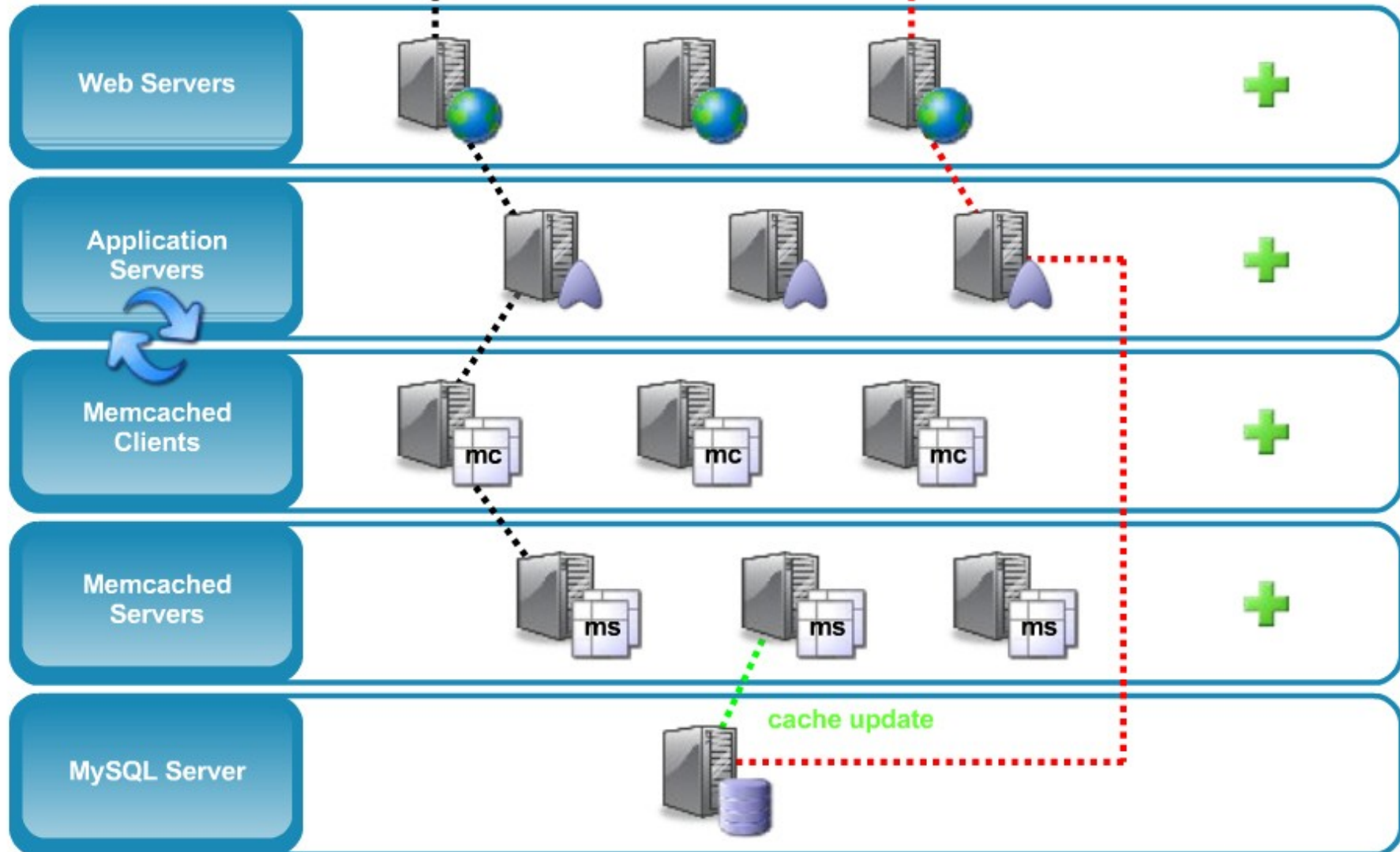
- Perlbal - <http://www.danga.com/perlbal/>
  - ◆ 多個網頁伺服器的負載平衡
  - ◆ Load balancer
- MogileFS - <http://www.danga.com/mogilefs/>
  - ◆ 分散式檔案系統
  - ◆ Distributed File System for small files
  - ◆ 有公司認為 MogileFS 比起 Hadoop 適合拿來處理小檔案
- memcached - <http://memcached.org/>
  - ◆ 共享記憶體 ??
  - ◆ Share Memory
  - ◆ 把資料庫或經常讀取的部分，用記憶體快取 (Cache) 方式存放
- Moxi - <http://code.google.com/p/moxi/>
  - ◆ Memcache 的 PROXY
- More Resource:
  - ◆ <http://code.google.com/p/memcached/wiki/HowToLearnMoreScalability>
  - ◆ <http://www.slideshare.net/techdude/scalable-web-architectures-common-patterns-and-approaches>



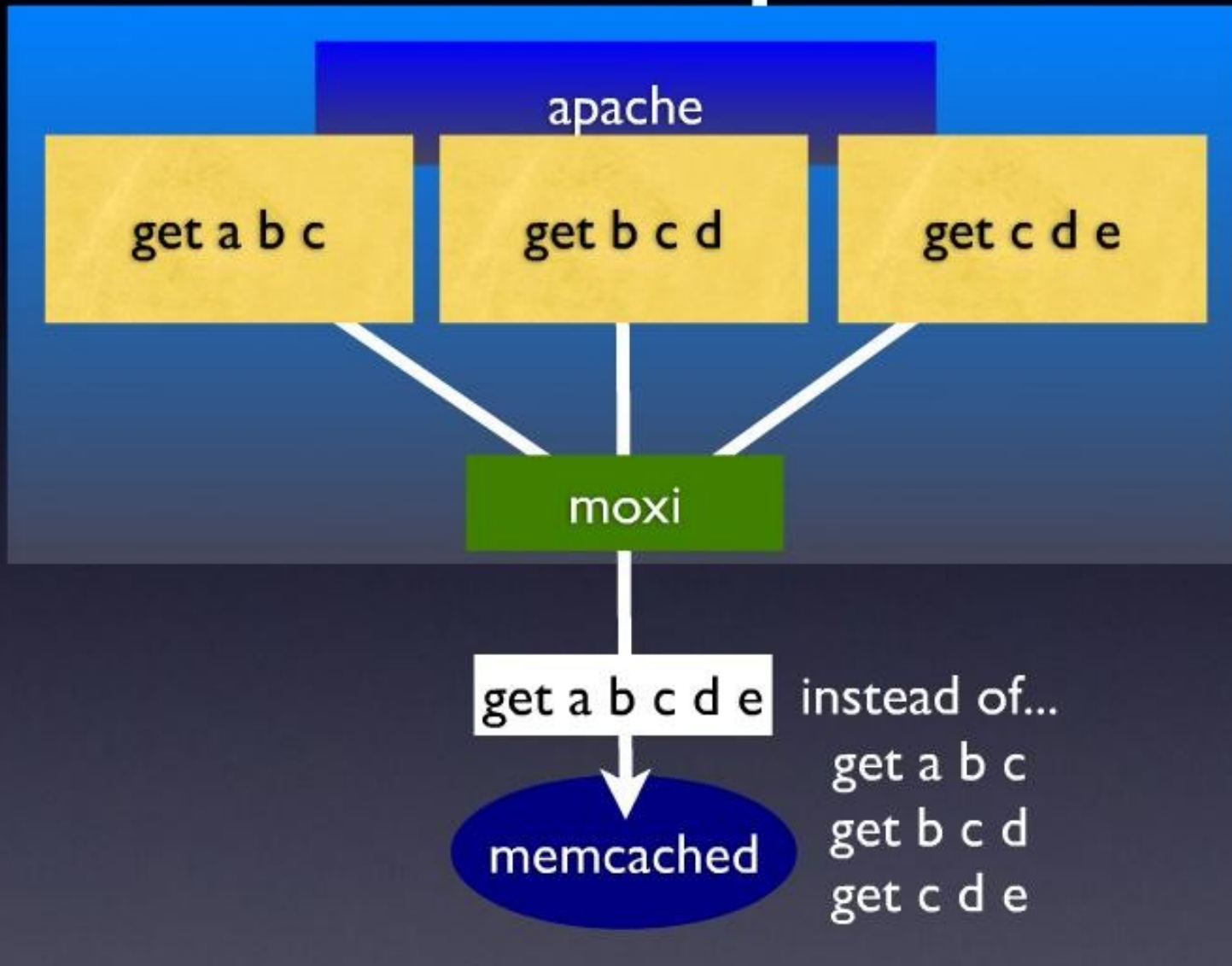
# Memcached & MySQL

read

write

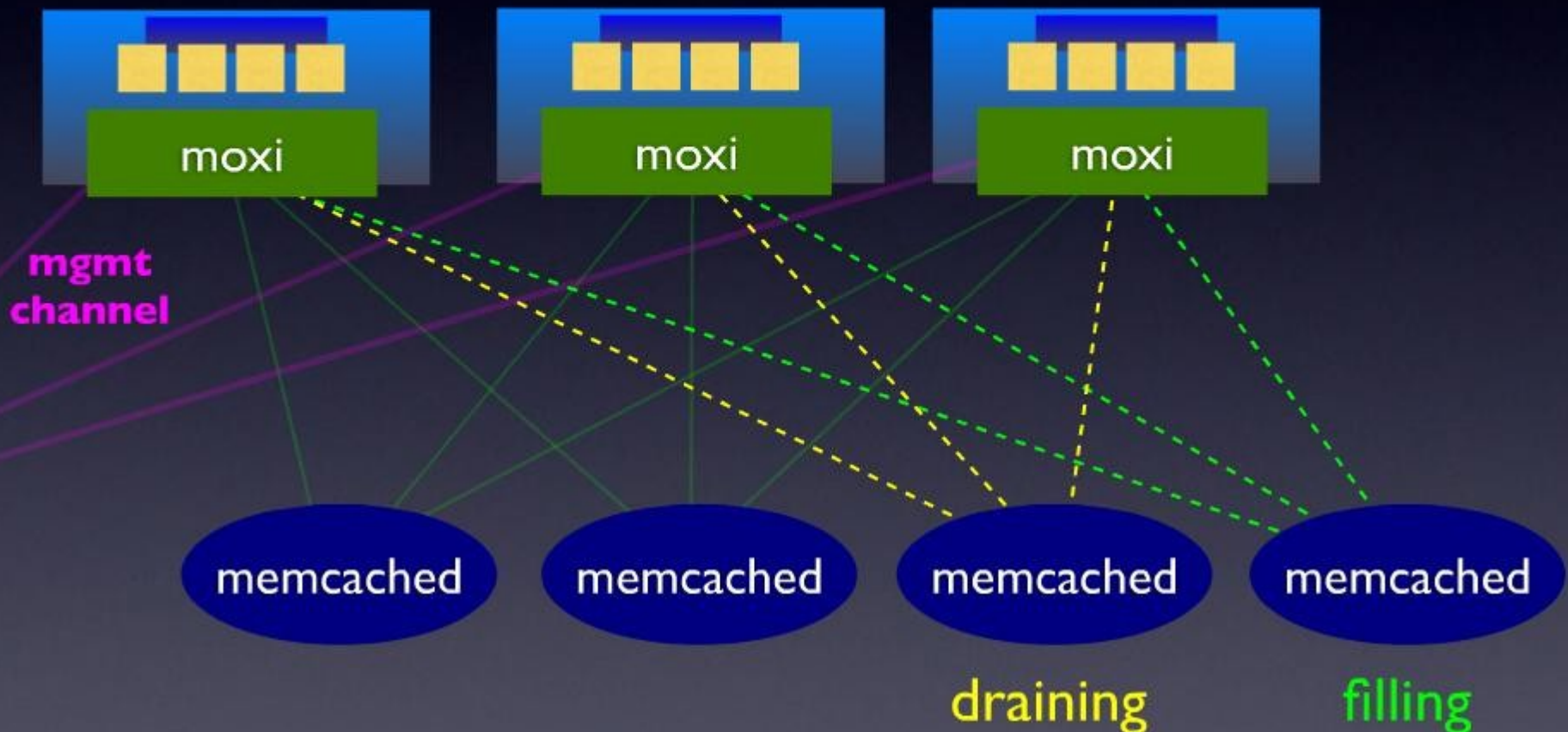


# GET de-duplication



# draining and filling

lazily migrate items from old server to new server





# HBase is ..

- HBase is a distributed **column-oriented database** built on top of HDFS.
- A distributed data store that can scale horizontally to 1,000s of commodity servers and **petabytes** of indexed storage.
- Designed to operate on top of the Hadoop distributed file system (**HDFS**) or Kosmos File System (**KFS**, aka Cloudstore) for scalability, fault tolerance, and high availability.
- Integrated into the Hadoop **map-reduce** platform and paradigm.

# Benefits

- Distributed storage
- Table-like in data structure
  - multi-dimensional map
- High scalability
- High availability
- High performance

# Who use HBase

- Adobe
  - 內部使用 (Structure data)
- Kalooga
  - 圖片搜尋引擎 <http://www.kalooga.com/>
- Meetup
  - 社群聚會網站 <http://www.meetup.com/>
- Streamy
  - Migrate from MySQL to Hbase <http://www.streamy.com/>
- Trend Micro
  - 雲端掃毒架構 <http://trendmicro.com/>
- Yahoo!
  - 儲存文件 fingerprint 避免重複 <http://www.yahoo.com/>
- More - <http://wiki.apache.org/hadoop/Hbase/PoweredBy>

# Backdrop

- Started toward by Chad Walters and Jim
- 2006.11
  - Google releases paper on **BigTable**
- 2007.2
  - Initial HBase prototype created as Hadoop contrib.
- 2007.10
  - First useable HBase
- 2008.1
  - Hadoop become Apache top-level project and HBase becomes subproject
- 2008.10~
  - HBase 0.18, 0.19 released

# HBase Is Not ...

- Tables have **one primary index**, the *row key*.
- **No join operators.**
- Scans and queries can select a subset of available columns, perhaps by using a wildcard.
- There are three types of lookups:
  - Fast lookup using row key and optional timestamp.
  - Full table scan
  - Range scan from region start to end.

## HBase Is Not ... (2)

- Limited atomicity and transaction support.
  - HBase supports **multiple batched mutations of single rows** only.
  - Data is unstructured and untyped.
- No accessed or manipulated via SQL.
  - Programmatic access via Java, REST, or **Thrift APIs**.
  - Scripting via JRuby.

# Why Bigtable?

- Performance of RDBMS system is good for transaction processing but for very large scale analytic processing, the solutions are commercial, expensive, and specialized.
- Very large scale analytic processing
  - Big queries – typically range or table scans.
  - Big databases (100s of TB)

## Why Bigtable? (2)

- Map reduce on Bigtable with optionally Cascading on top to support some relational algebras may be a cost effective solution.
- Sharding is not a solution to scale open source RDBMS platforms
  - Application specific
  - Labor intensive (re)partitioning



# Why HBase ?

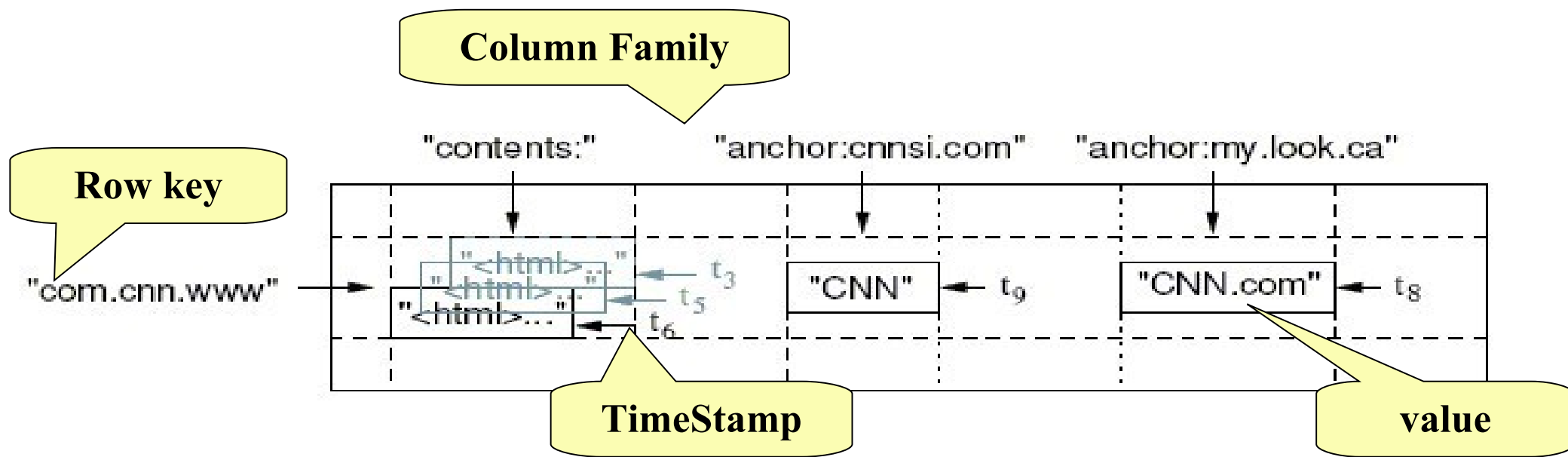
- HBase is a Bigtable clone.
- It is open source
- It has a good community and promise for the future
- It is developed on top of and has good integration for the Hadoop platform, if you are using Hadoop already.
- It has a Cascading connector.

# HBase benefits than RDBMS

- *No real indexes*
- *Automatic partitioning*
- *Scale linearly and automatically with new nodes*
- *Commodity hardware*
- *Fault tolerance*
- *Batch processing*

# Data Model

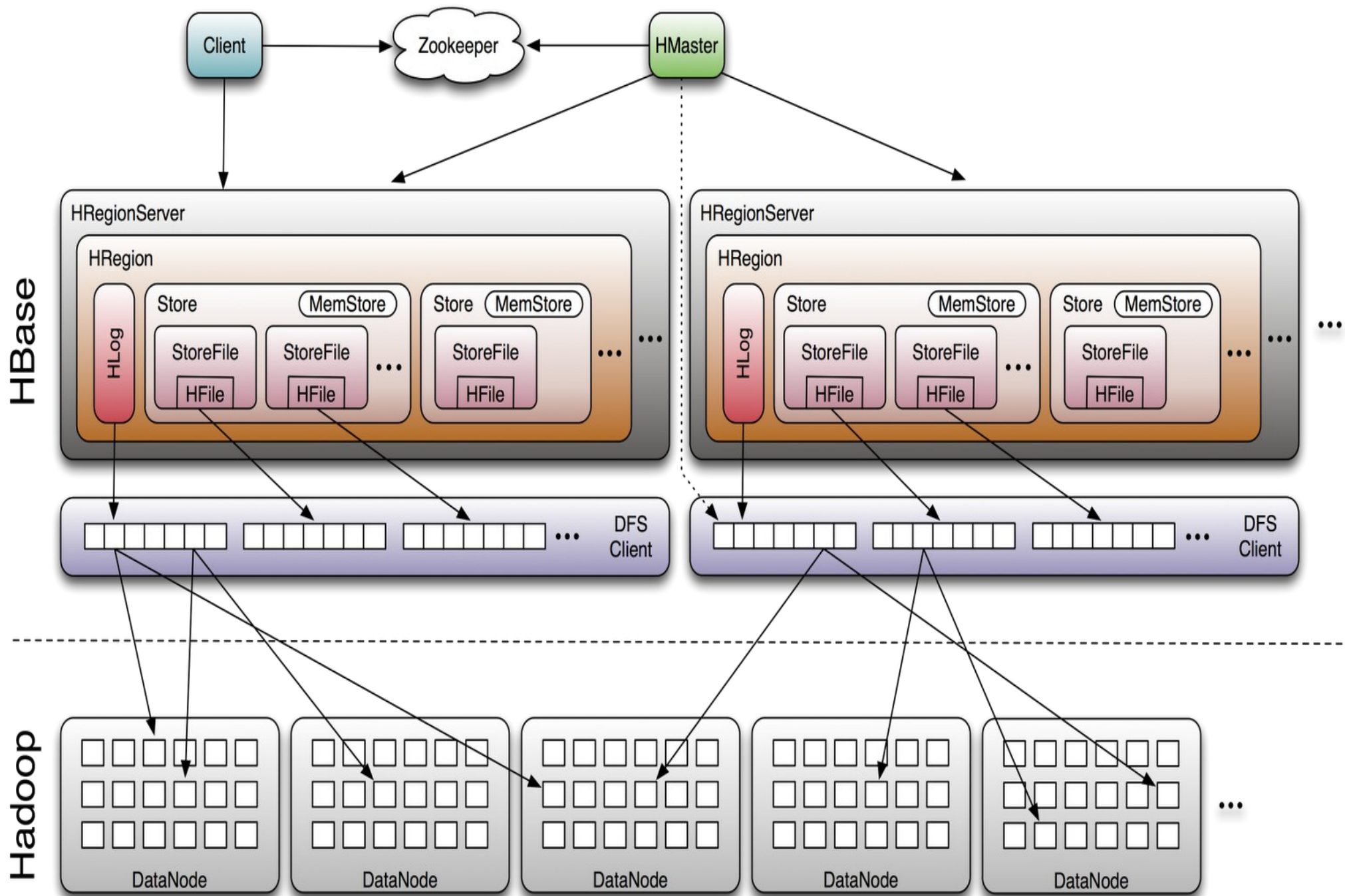
- Tables are sorted by **Row**
- Table schema only define it's *column families* .
  - Each family consists of any number of columns
  - Each column consists of any number of versions
  - Columns only exist when inserted, NULLs are free.
  - Columns within a family are sorted and stored together
- Everything except table names are byte[]
- **(Row, Family: Column, Timestamp) → Value**



# Members

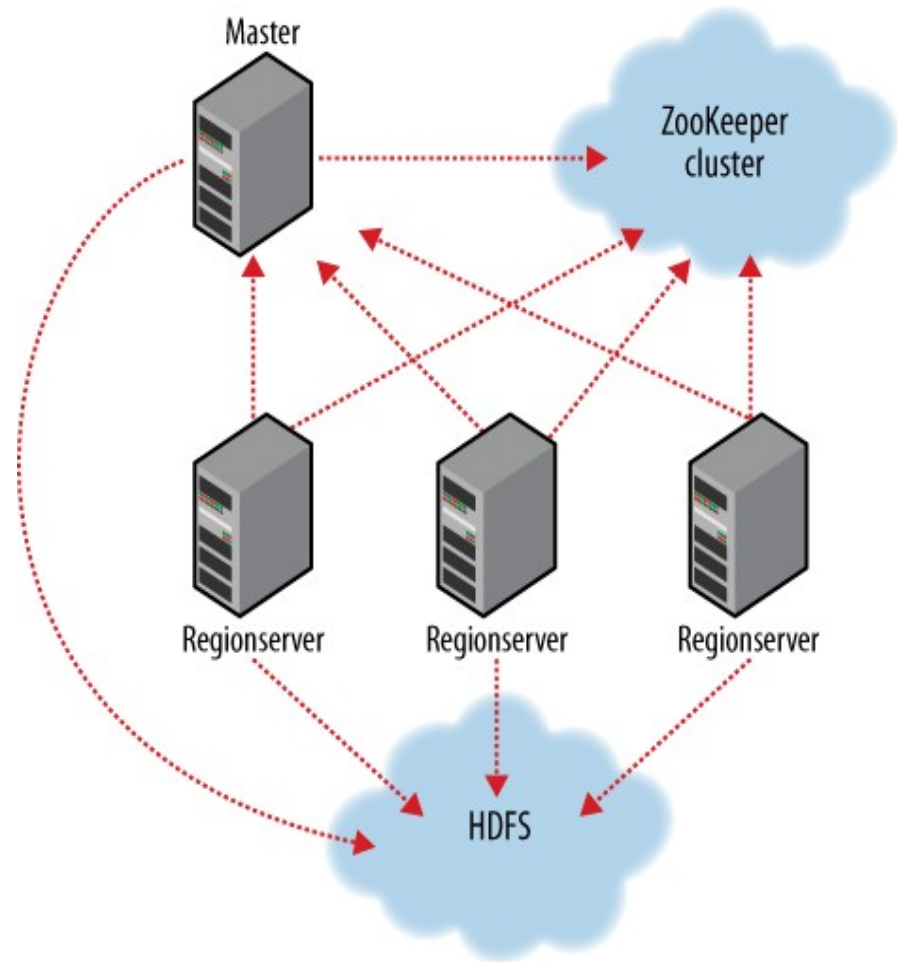
- *Master*
  - Responsible for monitoring region servers
  - Load balancing for regions
  - Redirect client to correct region servers
  - The current SPOF
- *regionserver slaves*
  - Serving requests(Write/Read/Scan) of Client
  - Send HeartBeat to Master
  - Throughput and Region numbers are scalable by region servers

# Architecture



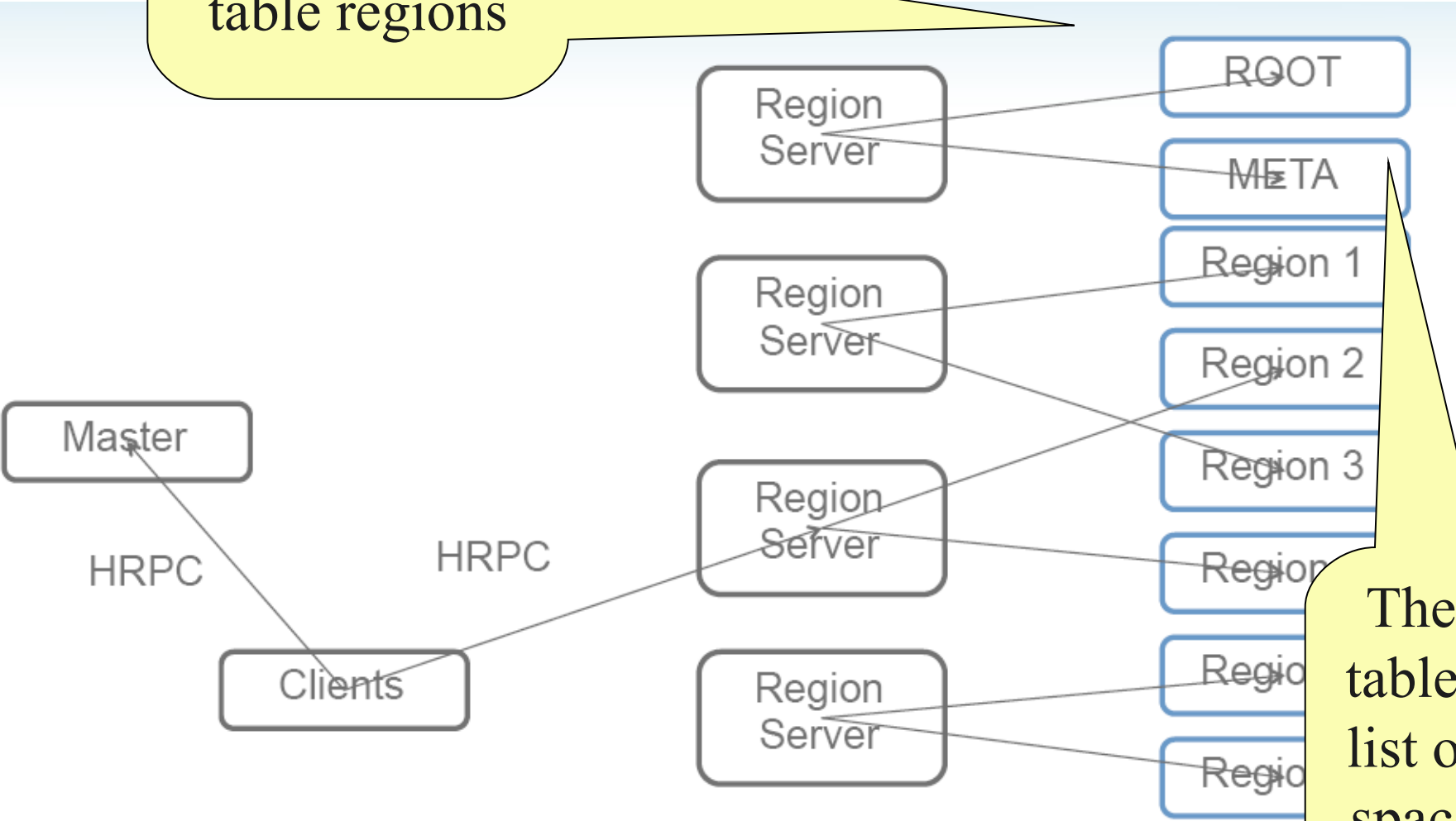
# ZooKeeper

- HBase depends on ZooKeeper (Chapter 13) and by default it manages a ZooKeeper instance as the authority on cluster state



# Operation

The `-ROOT-` table holds the list of `.META.` table regions



The `.META.` table holds the list of all user-space regions.