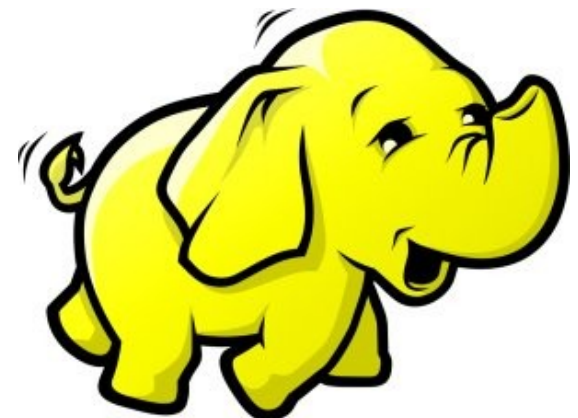# Hadoop 相關計畫
Hadoop Ecosystem

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^{TM}, 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^{TM} 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

- You can also find example code at
  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
```
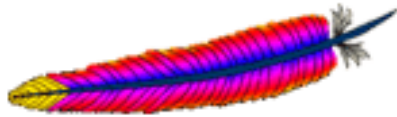
There are serveral Hadoop subprojects

- 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

Pig	Chukwa	Hive	HBase
MapReduce		HDFS	ZooKeeper
Hadoop Core (Hadoop Common)		Avro	

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 ...