Apache Hadoop

open-source software for reliable, scalable, distributed computing.

What is Apache Hadoop ?

- Hadoop Common

  The common utilities that support the other Hadoop modules.

- Hadoop Distributed File System (HDFS™)

  A distributed file system that provides high-throughput access to application data.

- Hadoop YARN

  A framework for job scheduling and cluster resource management.

- Hadoop MapReduce

  A YARN-based system for parallel processing of large data sets.

巨量資料

We are Big Data

I'm a small server

思 考 時 間

- 巨量資料帶給我們什麼？
- 我們該如何存放這些資料？
- 我們該如何處理這些資料？

想 像 時 間

Like a

想 像 時 間

Like

思 考 時 間

- 誰在用巨量資料？
- 什麼狀況可以使用巨量資料來處理？
- 巨量資料跟 Apache Hadoop 的關係？

Big Data 的起源 Google 三篇論文

- The Google File System
  http://research.google.com/archive/gfs.html

- MapReduce: Simplified Data Processing on Large Clusters
  http://research.google.com/archive/mapreduce.html

- Bigtable: A Distributed Storage System for Structured Data
  http://research.google.com/archive/bigtable.html

The Google File System Architecture

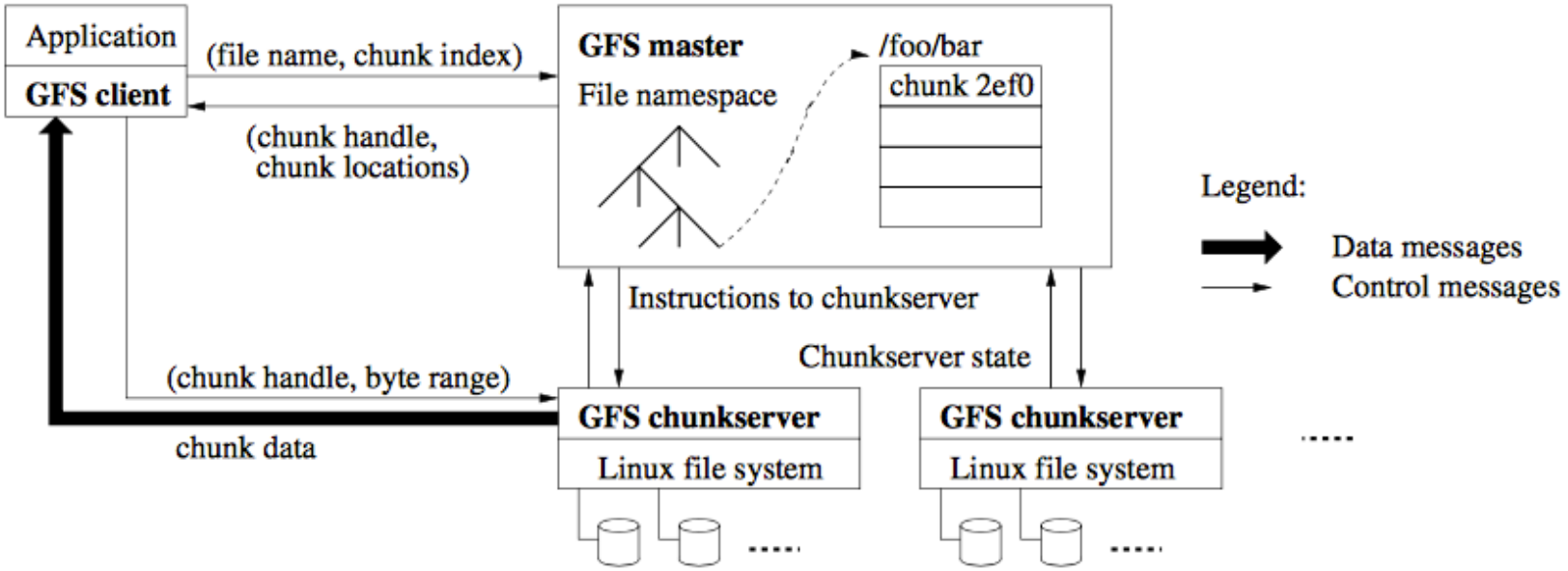

Figure 1: GFS Architecture

MapReduce Execution Overview

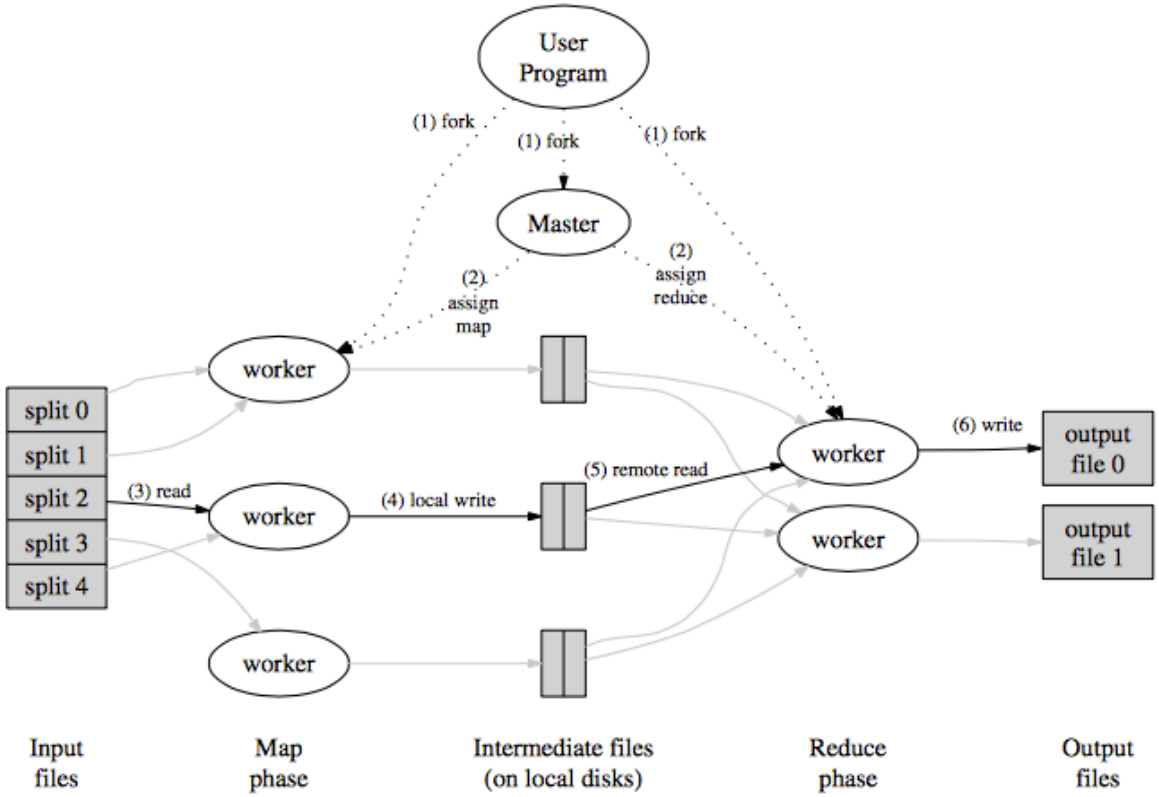

Figure 1: Execution overview

Apache Hadoop 環境建置

- Install VirtualBox
- Import Cloudera Quick Start VM
- Hadoop 基本操作指令