實作課程 Lesson 03

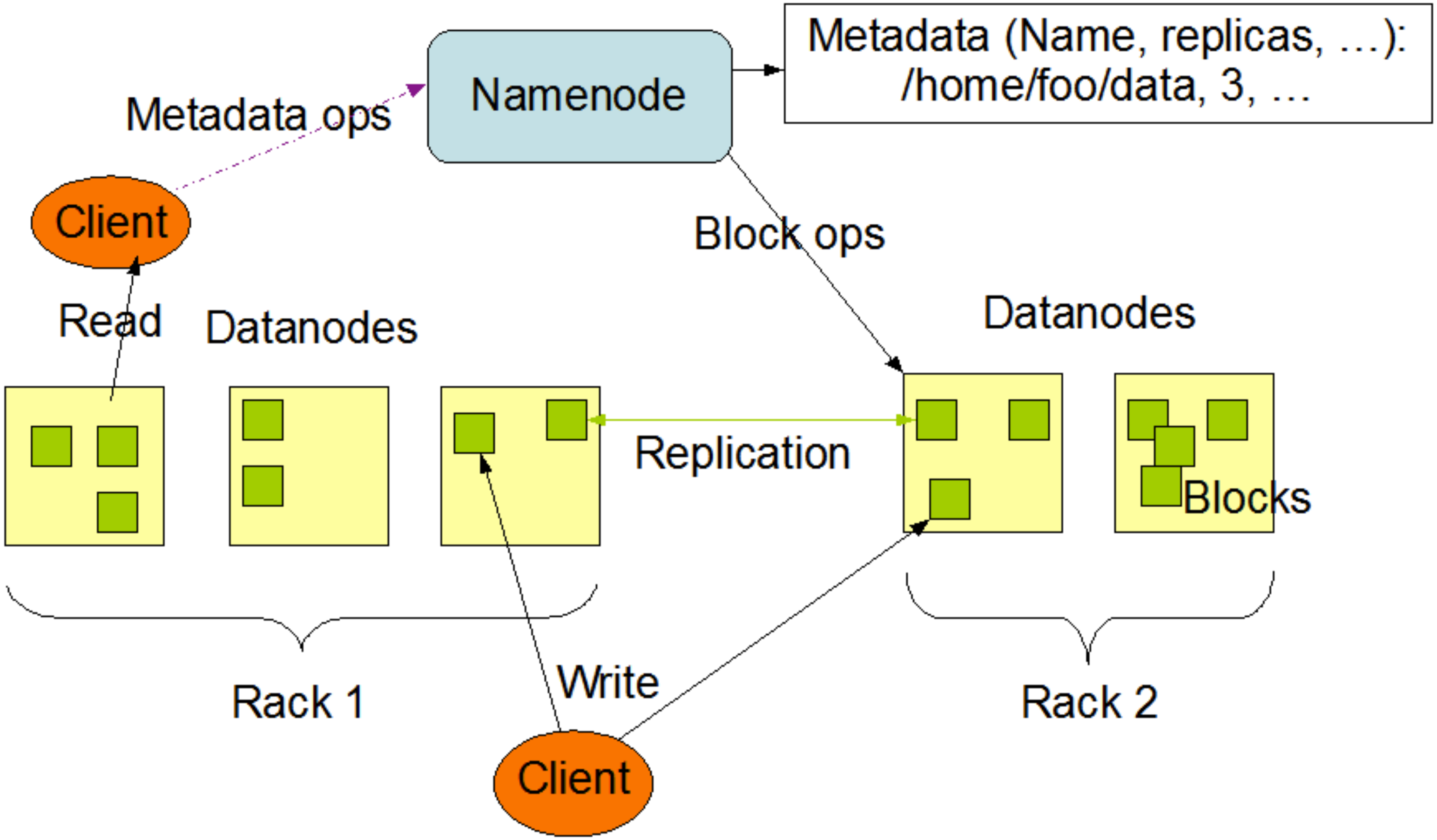
蔡秉文

Cookie Tsai

You Will Learn

1. 認識 MapReduce
2. 認識 Apache Maven
3. 如何使用 Apache Maven 打包 Jar
4. 如何使用 Hadoop Streaming
5. 如何使用 Hadoop 執行 MapReduce

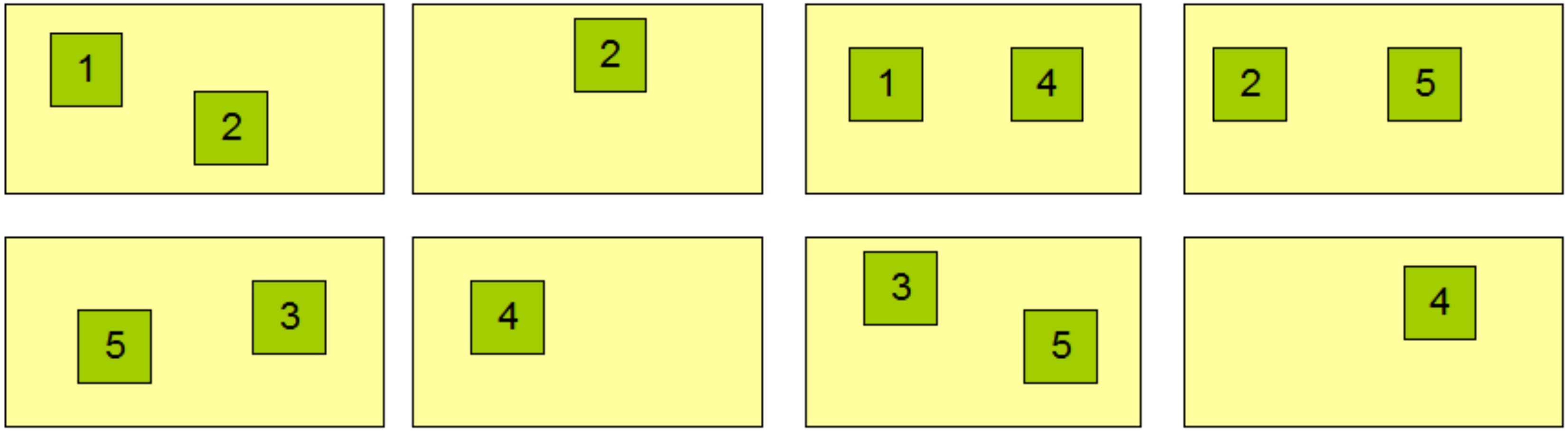
HDFS Architecture



Block Replication

Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …

Datanodes



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

MapReduce Execution Overview

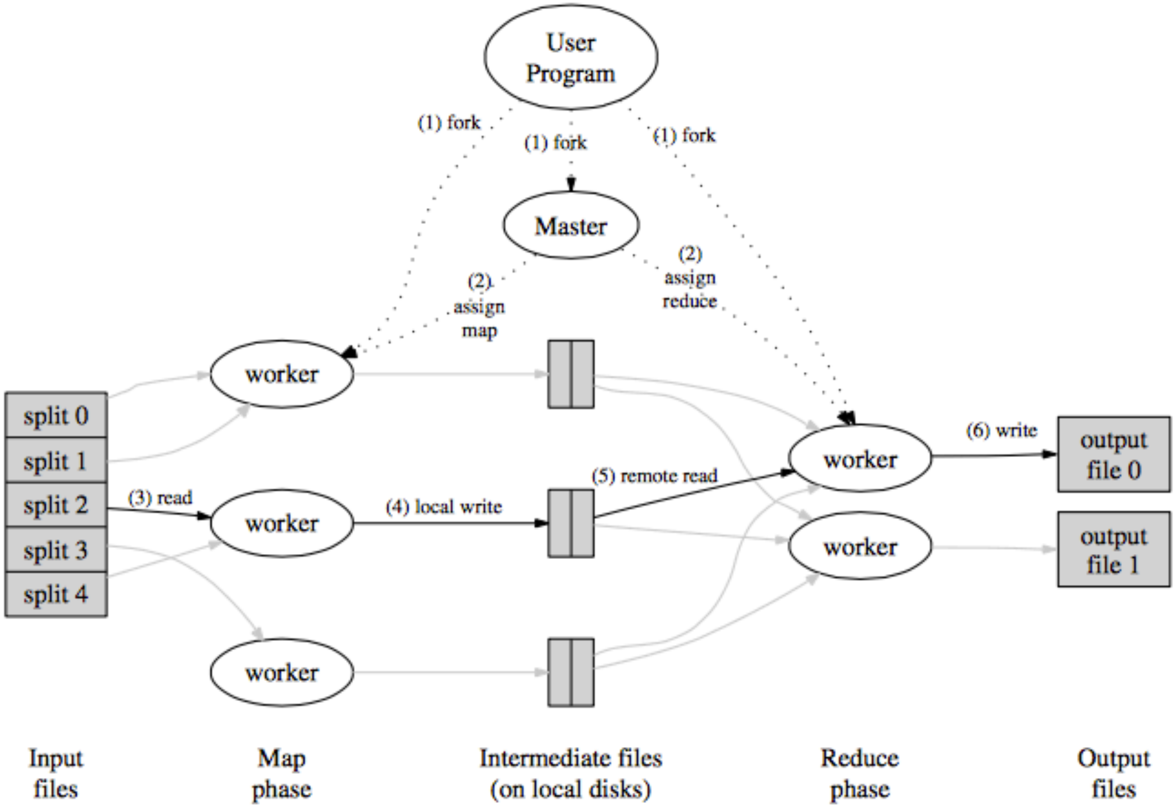


Figure 1: Execution overview

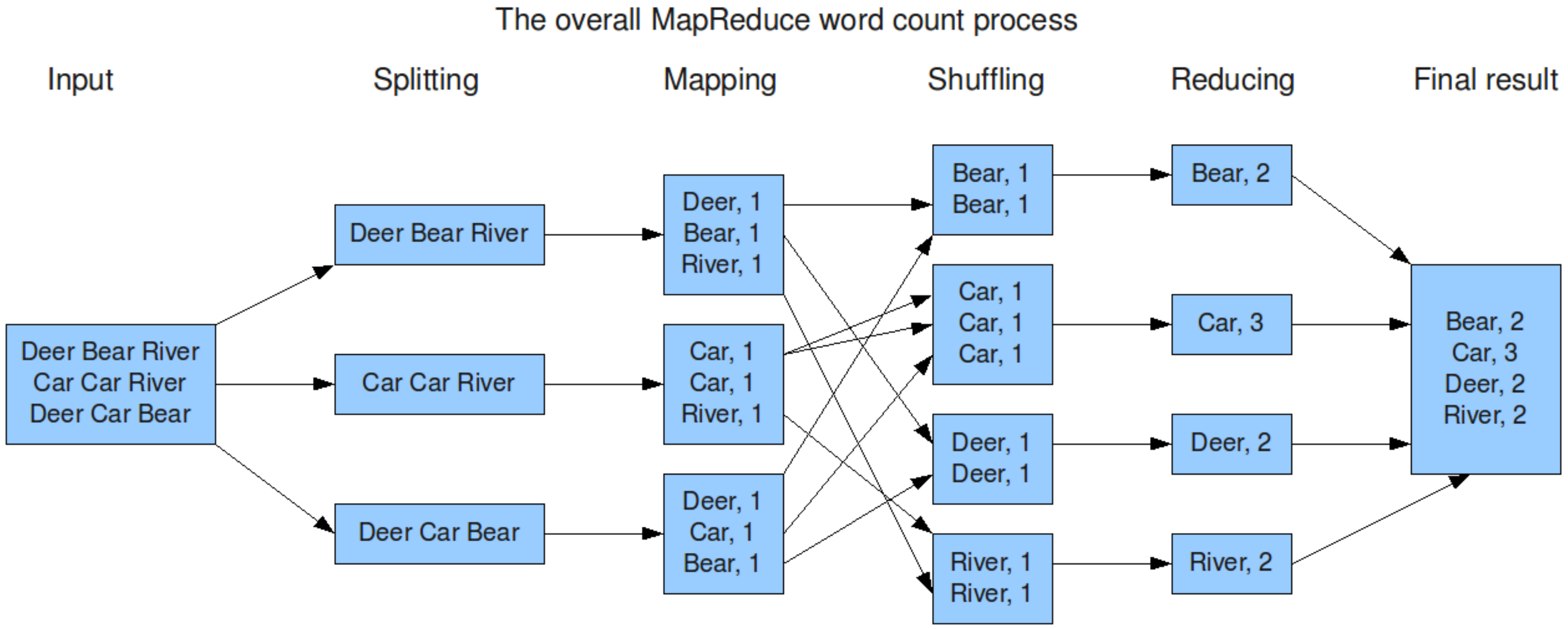
MapReduce Execution Overview

1. 將要執行的程式碼，複製到 Master 與每一臺 Worker 中
2. 由 Master 決定 Map 與 Reduce，分別由哪些 Worker 執行
3. 將資料區塊分配到執行 Map 的 Worker 中進行 Map
4. 將 Map 後的結果存入 Worker 機器的本地磁碟
5. 執行 Reduce 程式的 Worker 機器，遠端讀取每一份 Map 結果，進行彙整與排序，同時執行 Reduce
6. 將使用者需要的運算結果輸出至指定位置

What is MapReduce?

- MapReduce
  - 是一種軟體框架(software framework)
  - 能為大量資料做平行運算處理
  - 此框架的功能概念主要是映射(Map)和化簡(Reduce)
  - 實作上可用C++、JAVA或其他程式語言來達成
- Map
  - 從主節點(master node)輸入一組input，此input是一組key/value，將這組輸入切分成好幾個小的子部分，分散到各個工作節點(worker nodes)去做運算
- Reduce
  - 主節點(master node)收回處理完的子部分，將子部分重新組合產生輸出

What is MapReduce?



Make A MapReduce

- Getting Start with Maven
  - http://tsai-cookie.blogspot.tw/2016/03/getting-started-with-maven.html
- A Simple MapReduce Sample
  - https://github.com/CookieTsai/MapReduce

Reference

- MapReduce Tutorial
  - https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- Hadoop Streaming
  - https://hadoop.apache.org/docs/current/hadoop-streaming/HadoopStreaming.html
- 宅學習 – MapReduce, Hadoop
  - http://sls.weco.net/CollectiveNote20/MR

Thank you for your listening