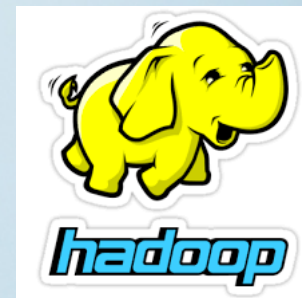




大數據資料處理 – Hadoop MapReduce Programming

Hadoop 介紹



授課教授：王經篤 (Jing-Doo Wang)

Email: jdwang@asia.edu.tw

亞洲大學 資訊工程系 (CSIE, Asia University)

課程大綱 (Outline)

- Hadoop History
- HDFS - Hadoop Distributed File System (HDFS)
- Map and Reduce Programming Model
- Hadoop Ecosystem

The History of Hadoop

Reference: <https://medium.com/@markobonaci/the-history-of-hadoop-68984a11704> (by [Marko Bonaci](#))



Doug Cutting ✓

@cutting

Founder of Lucene, Nutch, Hadoop & Avro.

📍 California, USA

🌐 blog.lucene.com

From:

https://twitter.com/cutting?ref_src=twsrc%5Etfw&ref_url=https%3A%2F%2Fmedium.com%2Fmedia%2Fce9478b94bd8dfa7835051b6551e3e91%3FpostId%3D68984a11704

新聞

臺灣資料工程協會即將成立，大數據平臺Hadoop之父

大數據分析不只需要資料科學家或資料分析師，實務上，國外常見還需要資料工程師和資料架構師的角色，培養這兩類人才正是臺灣資料工程協會的成立目的。

文/ 王宏仁 | 2017-07-11 發表

✓ 讚 4.4 萬 按讚加入iThome粉絲團

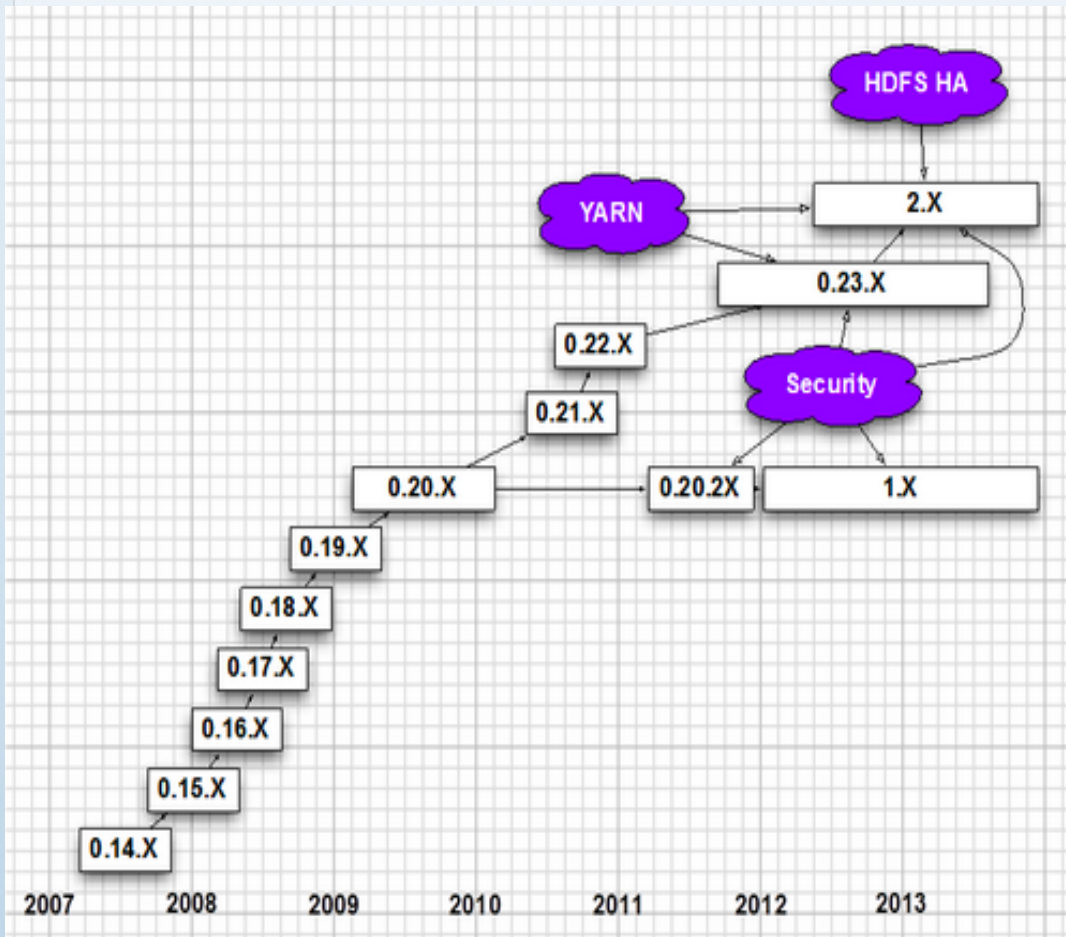
👍 讚 981 分享



From: <http://www.ithome.com.tw/news/115498>

Hadoop 版本演進史

07 July, 2017, Hadoop 3.0.0-alpha4



Source: P.6, [Running YARN at Scale](#)

Apache > Hadoop >

Search with Apache Solr

Last Published: 08/06/2017 16:25:10

Top **Wiki**

About

- Welcome
- Releases
- Download
- Release Notes
- Release Versioning
- Mailing Lists
- Issue Tracking
- Who We Are?
- Who Uses Hadoop?
- Buy Stuff
- Sponsorship
- Thanks
- Privacy Policy
- Bylaws
- Committer criteria
- License

Documentation

Apache Hadoop Releases

Download

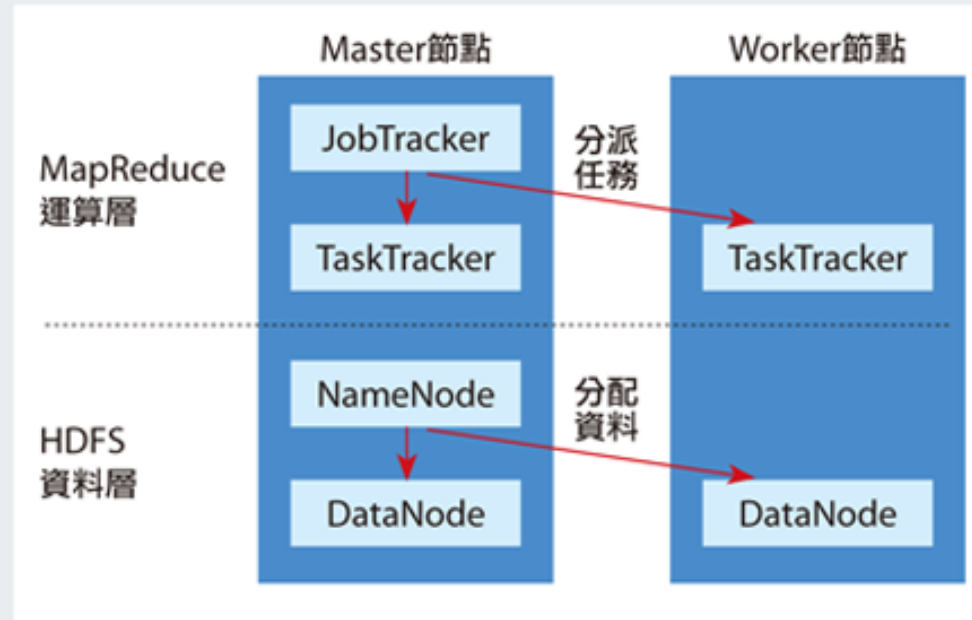
Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-256.

Version	Release Date	Tarball	GPG	SHA-256
3.0.0-alpha4	07 July, 2017	source	signature	checksum file
		binary	signature	checksum file
2.8.1	08 June, 2017	source	signature	0748C0E2 519382F2..
		binary	signature	B5BE5275 78EF2C85..
2.7.4	04 August, 2017	source	signature	D52B8CE8 446F4C10..
		binary	signature	8F791BFC FA5BB7C7..
2.6.5	08 October, 2016	source	signature	3A843F18 73D9951A..
		binary	signature	001AD18D 4B6D0FE5..

Hadoop V1

Hadoop 架構的伺服器角色分工

Hadoop運算叢集中的伺服器依用途分成Master節點和Worker節點。Master節點中安裝了JobTracker、NameNode、TaskTracker和DataNode程式，但Worker節點只安裝TaskTracker和DataNode。



From: <http://www.ithome.com.tw/node/73978>

Single Use System

Batch Apps

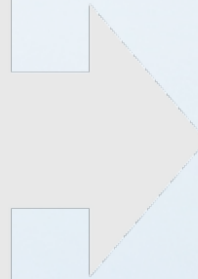
HADOOP 1.0

MapReduce

(cluster resource management
& data processing)

HDFS

(redundant, reliable storage)



Multi Use Data Platform

Batch, Interactive, Online, Streaming, ...

HADOOP 2.0

MapReduce

(batch)

Tez

(interactive)

Others

(varied)

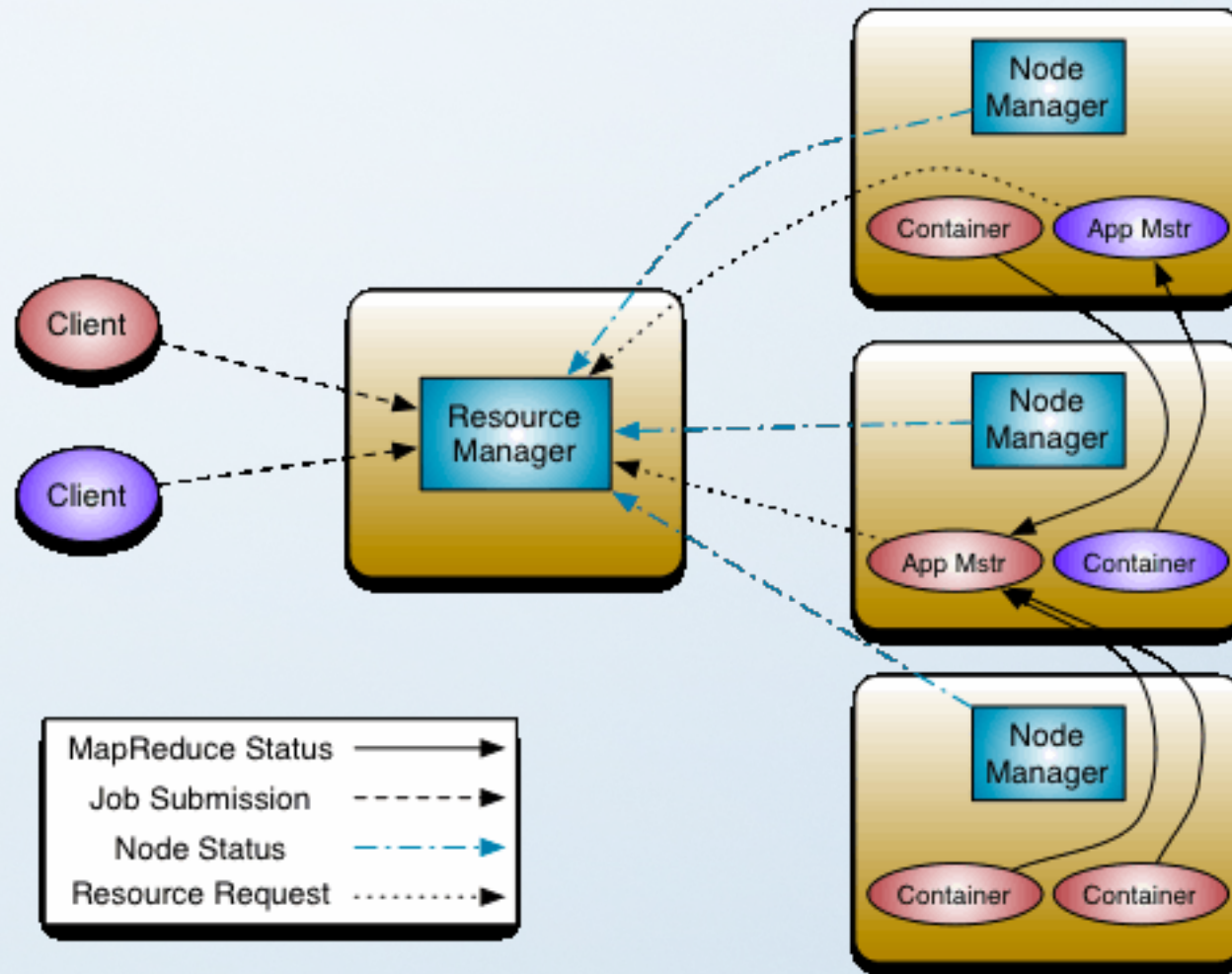
YARN

(operating system: cluster resource management)

HDFS2

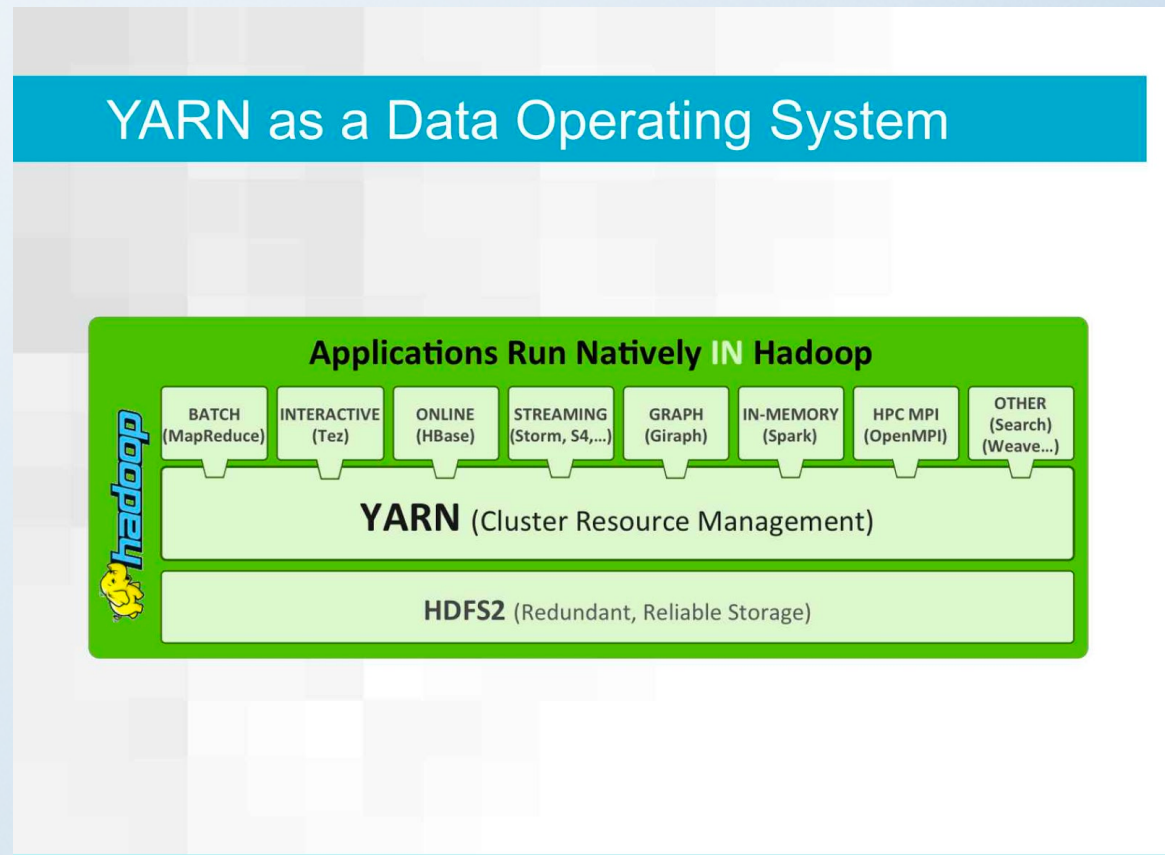
(redundant, reliable storage)

Resource Manager vs. Node Manager (Hadoop V2, MRv2 (Yarn))

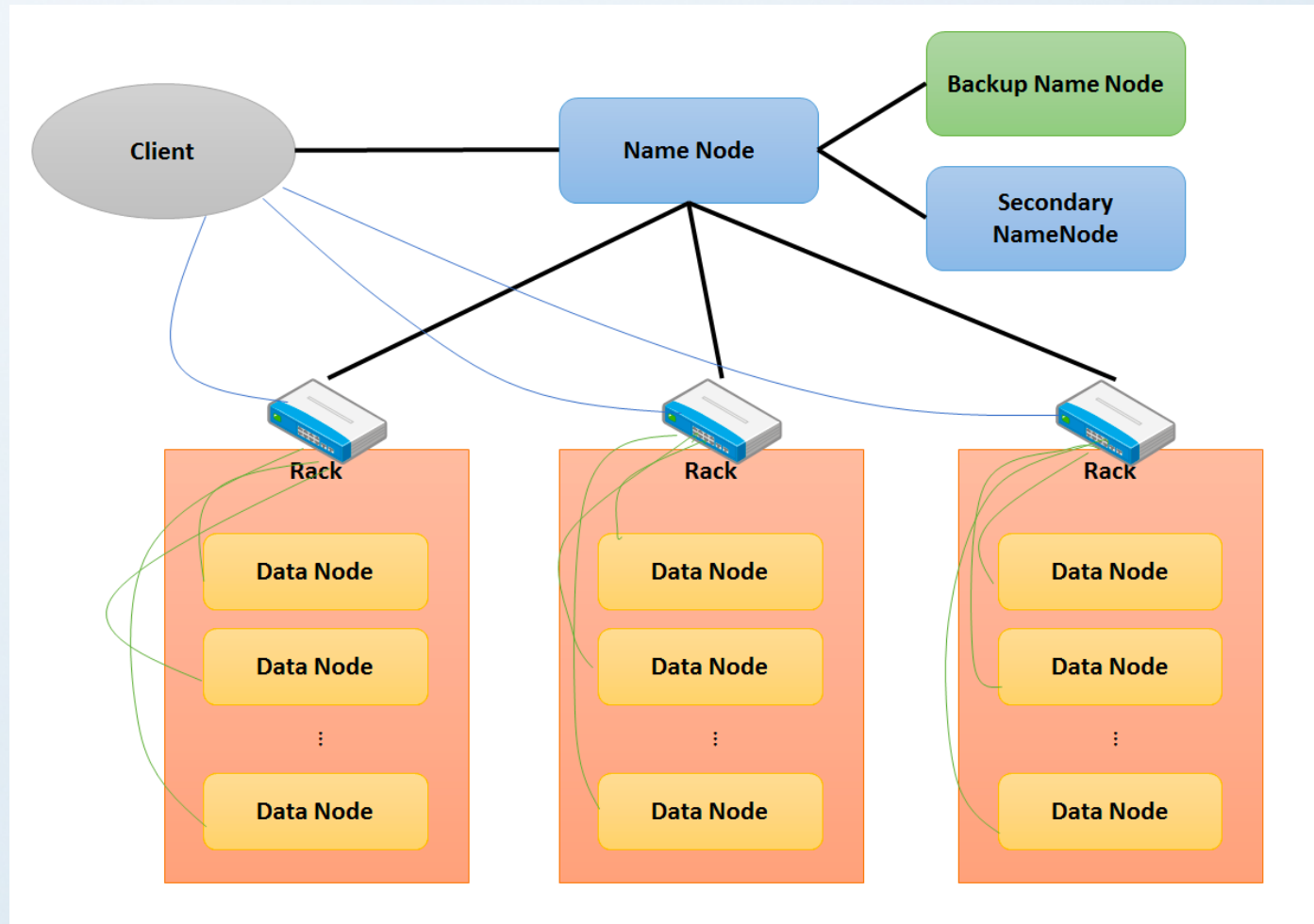


YARN (Yet Another Resource Negotiator)

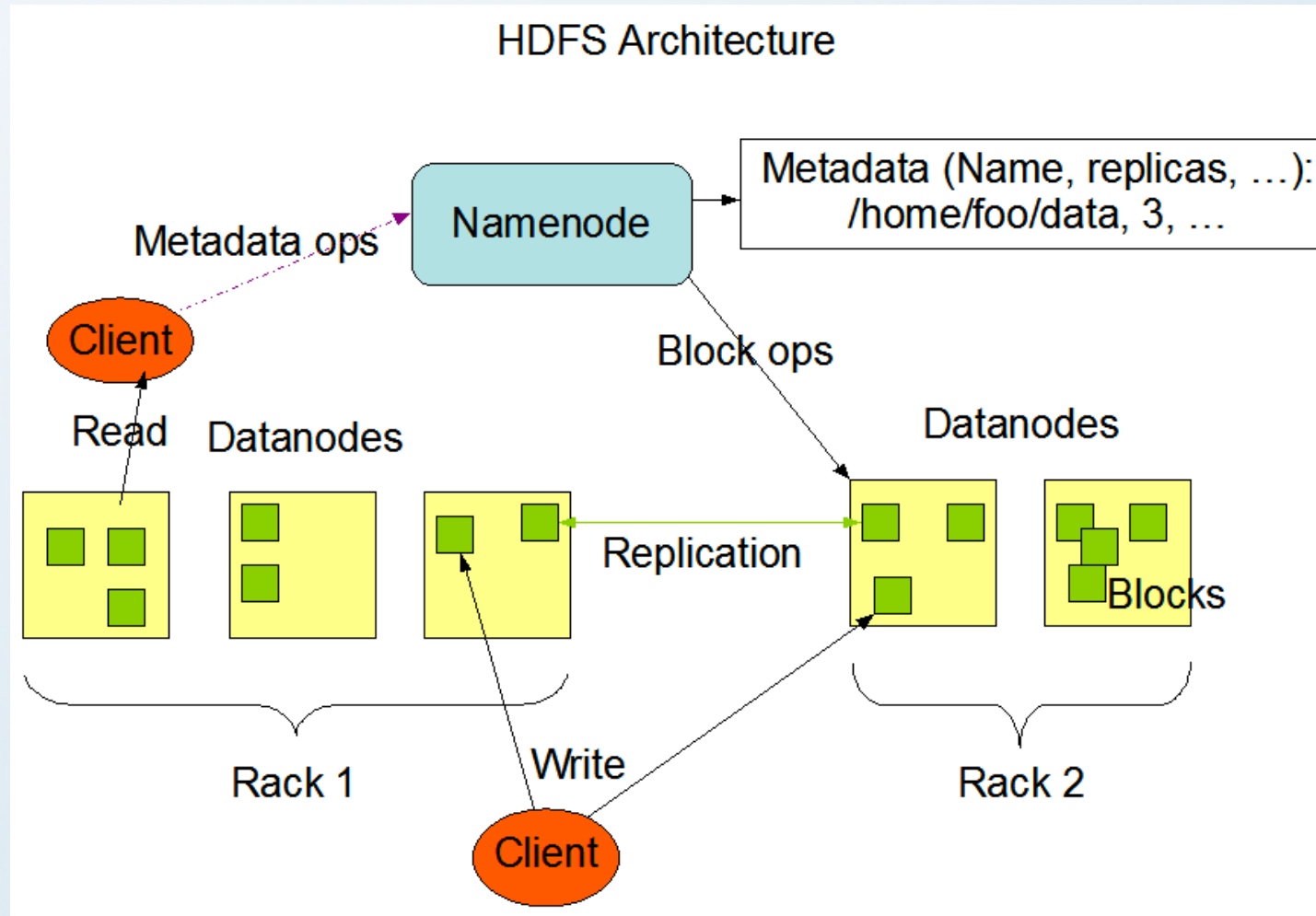
- YARN is the framework responsible for **assigning computational resources** for application execution.



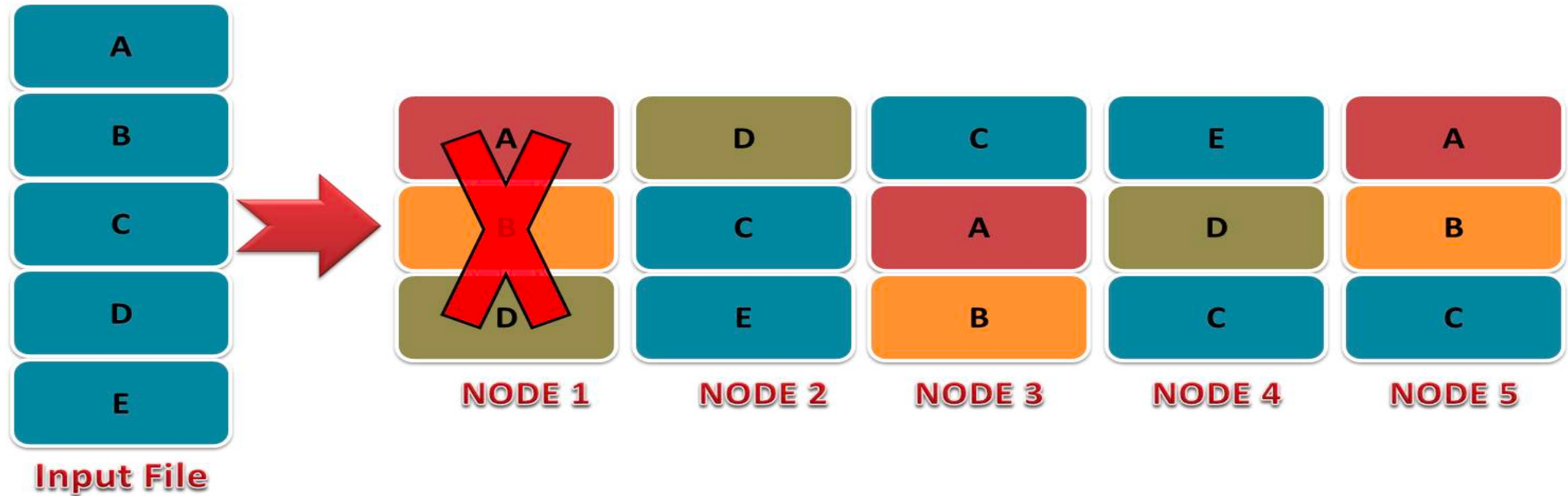
HDFS - Hadoop Distributed File System



HDFS- Fault Tolerance



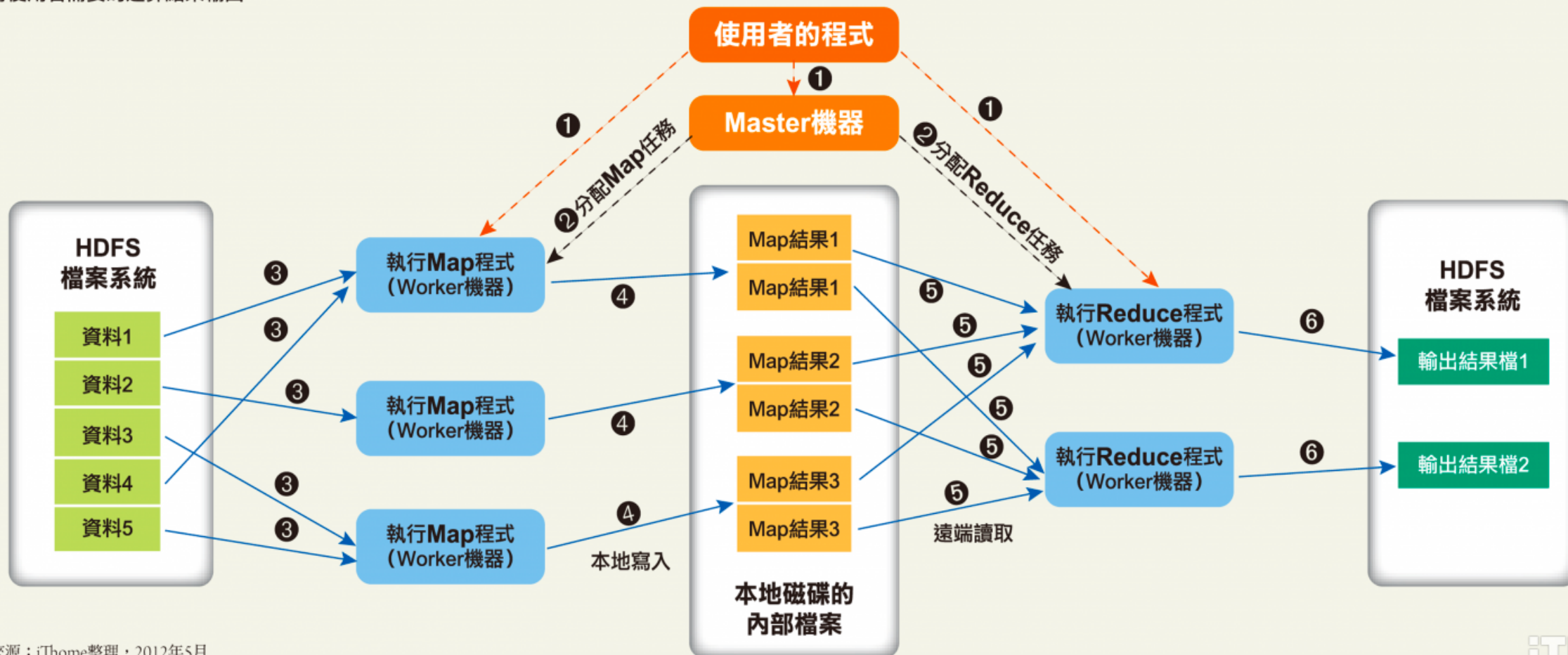
HDFS Feature – Fault Tolerance



MapReduce 執行示意圖

MapReduce是Hadoop分散式運算的關鍵技術，將要執行的問題，拆解成Map和Reduce的方式來執行，以達到分散運算的效果。例如要搜尋哪些網頁中有「iThome」這個字，可以先用Map程式，來計算出所有網頁中，每一個字的位置。再使用Reduce程式，在每一個字的清單中，篩選出「iThome」這個字，所對應的網頁網址。MapReduce程式的執行過程如下：

- 1 將要執行的MapReduce程式複製到Master與每一臺Worker機器中。
- 2 Master決定Map程式與Reduce程式，分別由哪些Worker機器執行。
- 3 將所有的資料區塊，分配到執行Map程式的Worker機器中進行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程式。
- 6 將使用者需要的運算結果輸出。



資料來源：iThome整理，2012年5月

iThome

Hadoop Ecosystem

