



進階課程

Hadoop + RDBMS

© TemplatesWise.com

Hadoop 0.20 + MySQL 5.5

王耀聰 陳威宇

Jazz@nchc.org.tw

waue@nchc.org.tw



財團法人國家實驗研究院

國家高速網路與計算中心

NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING



Hadoop 與 JDBC

JDBC：用於執行 SQL 的 Java API，目的在於無論機器改變、或資料庫改變，程式碼都可以不用改變

支援 MySQL, PostgreSQL, Oracle, SQL...

非預設就有，需額外針對資料庫形式下載對應的 Jar 檔

Hadoop 0.19 開始設計能與 JDBC 溝通的 API

*org.apache.hadoop.**mapred.lib.db** (0.19)*

*org.apache.hadoop.**mapreduce.lib.db** (0.20~0.21)*

DBConfiguration

DBConfiguration.configureDB (參數)

參數：

job	driverClass	dbUrl	userName	passwd
Hadoop 定義的 job	JDBC driver	資料庫 url	帳號	密碼
job	com.mysql.jdbc.Driver	jdbc:mysql://localhost/mydb	root	rootpasswd

DBInputFormat

從 DB 端讀取資料

DBRecordReader : 從 table 中讀取 bytes 記錄

DBInputSplit : 描述輸入 bytes 的範圍

setInput() : 設定初始輸入資料

DBWritable

DBWritable : 定義 key-value 格式

```
public void write(PreparedStatement  
statement) throws SQLException;
```

```
public void readFields(ResultSet resultSet)  
throws SQLException;
```

DBOutputFormat

對 DB 端寫入資料

`DBOutputFormat.setOutput()` 設置輸出資訊

Conclusions

Hadoop 可以支援 RDBMS 與 NoSQL

NoSQL 的 HBase 結構上與 Hadoop 設計相近，又是針對存取大資料量，彼此互相支援

透過 Java 的 JDBC ， Hadoop 也可以對 MySQL 等資料庫作存取動作