**Loader DataValidation**

**Version 1.0.0**

**目錄**

1. **介紹**

**1.1 DataValidation工具是什麼?**

**1.2 DataValidation 如何實作?**

**1.3 執行環境**

**2. 檔案說明**

**2.1 bin**

**2.2 conf**

**2.3 lib**

**3. 使用說明**

**3.1 解壓縮**

**3.2 修改設定檔**

**3.3 執行指令**

**4. 執行結果**

**4.1 欄位名稱、型態一致，資料也一致**

**4.2 欄位名稱、型態一致，資料不一致**

**4.3 欄位名稱、型態不一致**

**1. 介紹:**

**1.1 DataValidation 工具是什麼?**

DataValidation是一個可以將兩種資料庫內的資料表作比對資料的工具，

當使用者想驗證某兩個資料表的內容是否一致時，便可使用

DataValidation，來達到此目的的效果。

**1.2 DataValidation 如何實作?**

DataValidation 運用JDBC 這個API去連接各種資料庫，只需要擁有對應的

ODBC驅動，幾乎可以存取所有的資料庫。

當連接到資料庫後，便可以下sql的搜尋指令，依照primary key的大小，

由小到大讀取資料表的資料來做比對。

其中，我們以**來源端**資料作為比對的標準，將兩邊的資料表每次讀取

一筆row來做比對，比對完將不一致的資料放置在一個LIST內，

儲存不一致的資訊。等待全部資料比對完後，輸出相對應的資訊。

* 1. **執行環境**

DataValidation 需要運行在hadoop的cluster上面。

1. **檔案說明:**

將 DataValidation-1.0.0-package.tar.gz解壓縮後，

在DataValidation-1.0.0資料夾中，會有三個資料夾。

**2.1 bin**

存放Validate script檔，負責建立java程式所需的環境資料，

並用來執行此專案。

**2.2 conf**

存放sql.xml，DataValidation的設定檔範例。

**2.3 lib**

存放此專案所需要之函式庫，包含

-CLABS\_PHOENIX-4.5.2-1.clabs\_phoenix1.1.0.jar (Phoenix JDBC)

-mysql-connector-java-6.0.2.jar (Mysql JDBC)

-DataValidation-1.0.0.jar (此專案)

1. **使用說明:**

**3.1 解壓縮**

將DataValidation-1.0.0-package.tar.gz放置於執行環境下解壓縮，

得DataValidation-1.0.0資料夾。

**3.2 修改設定檔**

進入到DataValidation-1.0.0資料夾內的conf資料夾，修改

sql.xml範例檔案的值，只需修改值即可。

(a) validation.database.source.url 來源資料庫(mysql)

(b) validation.database.target.url 目標資料庫(phoenix)

(c) validation.input.table.name 來源的資料表名稱

(d) validation.output.table.name 目標的資料表名稱

(e) validation.database.primarykey 須提供來源primarykey欄位名稱

**3.3 執行指令**

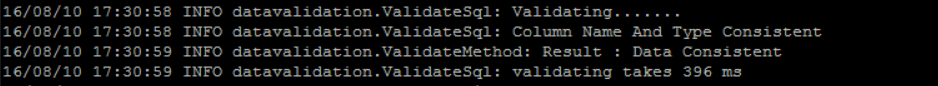
到DataValidation-1.0.0資料夾內，鍵下

Bash ./bin/validate ./conf/sql.xml指令執行

1. **執行結果:**

DataValidation執行結果我們主要分成三類，依照欄位名稱與型態的不一致做第一次的比對，比對成功後再執行內容資料的比對，再將比對完的結果輸出。

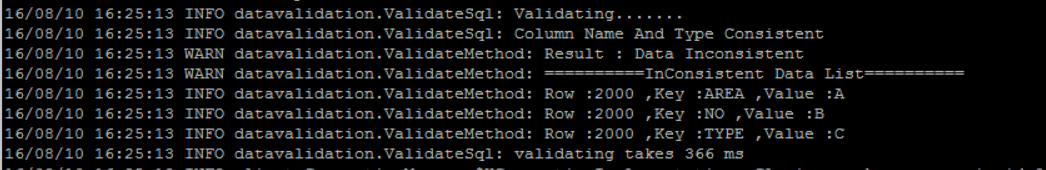
**4.1 欄位名稱、型態一致，資料也一致**



結果如上圖的資訊，可知道欄位名稱、型態驗證成功，資料也一致外，

最後會出現此次驗證所花的時間。

**4.2 欄位名稱、型態一致，資料不一致**



除了給予欄位名稱、型態驗證成功，資料不一致的資訊外，也會給予

不一致資料的的相關資訊。如上圖: Row:2000 ,Key:AREA ,Value:A

Row : 以primary key小到大的排序的第2000筆

Key : 此Row的欄位名稱AREA

Value : 此欄位AREA的值A與來源不一致

**4.3 欄位名稱、型態不一致**



欄位名稱、型態不一致時，資料當然也就不用繼續驗證下去!