

從關聯式資料到 HBase 資料的轉換

陳靖國*

朝陽科技大學資訊管理系
e-mail: jkchen@cyut.edu.tw

李維哲

朝陽科技大學資訊管理系
e-mail: s10514613@cyut.edu.tw

摘要

關聯式資料庫目前被廣為使用，雖然查詢所得的資料精確，但是複雜問題的查詢，合併表格需要較長的時間，不利於只需要查詢時間快但不講究資料精確的應用，而 Hadoop HBase 剛好有此特性。所以本文提出一個轉換方法，把關聯式資料庫資料轉換成 Hadoop HBase 的非關聯式相關資料。讓企業目前還在使用的關聯式資料庫資料可以繼續使用，而不必建新的 HBase 資料庫，節省企業成本。

關鍵詞：關聯式資料庫、Hadoop、HBase、資料轉換

Abstract

Relational database is currently widely used, although the query information accurate, but complex queries, the need to join the forms takes a long time, not only need to query the time fast but do not pay attention to the application of accurate data, and Hadoop HBase just have this feature. Therefore, this paper presents a conversion method that converts the relational database data into non-relational related data of Hadoop HBase. So that is still in use with the associated database information can also be re-use, saving business costs.

Keywords: Relational database, Hadoop, HBase, Data conversion

1. 前言

關聯式資料庫由於具有資料一致性與最少重複值的特性，且多數的關聯式資料庫管理系統具備有圖形化界面，所以使用和維護起來相對較為容易，也是目前被廣為使用的資料庫[1]。當要解決複雜問題需要大量 JOIN 資料表時，處理的速度就會變慢。

以 Google 的 MapReduce[8]和 HDFS 為基礎的 Hadoop，可支援分散式應用[3]。Hadoop 平台包括 Hadoop 核心、MapReduce、HDFS 以及一些相關項目，有 Hive[5]和 HBase[4]等。HBase 是個非關聯式分散式資料庫 (NoSQL)，提供類似 Google 公司 BigTable[6]

的服務，應用於多個資料驅動型網站，例如 Facebook 的訊息平台。

關聯式資料庫查詢資料的效率比 HBase 低落，因為 HBase 是將多個資料表有相關的資料都存放在同一個資料表裡，這樣在同時用到這些資料時可以減少多個表格之間 JOIN 所需要的時間。如果一個關聯式資料庫在某種應用上常常要 JOIN 多個資料表，則將其改建為 HBase 可以改善資料處理效率。所以本文依據 HBase 的資料模型，提出一個方法將目前現有的關聯式資料表轉換到 HBase 的資料表，加快大量資料的查詢速度[10]。

2. 相關文獻

2.1 關聯式資料庫

關聯式資料庫模型是一種由二維表格所構成的資料結構，由關聯式綱要 (Relation Schema) 和關聯表實例 (Relation Instance) 所構成。其中關聯式綱要描述關聯表名稱、屬性名稱和其定義域，其中定義域是指屬性值可接受的資料或資料型態；關聯表實例則是指某一時間點儲存在關聯表的資料，以二維表格所構成[1]。以圖 1 為例，這張關聯表的名稱為 Students，有 sid、name、address、tel、birthday 五個屬性，而屬性所屬的 Integer、string 和 date 皆屬於定義域。

關聯式綱要 (Relation Schema)	Students				
	sid: Integer	name: string	address: string	tel: string	birthday: date
關聯表實例 (Relation Instance)	S001	江小魚	中和豐平路1000號	02-22222222	1978/2/2
	S002	劉得華	桃園市三民路1000號	03-33333333	1982/3/3
	S003	郭富成	台中市中港路三段500號	04-44444444	1978/5/5
	S004	張學友	高雄市中區維多利亞路1000號	05-55555555	1979/6/6

圖 1. 關聯式資料模型的資料結構

關聯式資料庫模型除了上述所談論到的資料結構之外，完整性限制條件 (Integrity Constraints) 也是資料庫設計的一部分，包括鍵限制條件、定義域限制條件、實體完整性限制條件和參考完整性限制條件。鍵限制條件是指關聯表一定要擁有一個唯一和最小的主鍵，常見的主鍵限制條件包括用來識別值組是唯一的主鍵和參考其他關聯表的外來鍵。定義域限制條件是指在關聯表的屬性值一定要符合

定義域的單元值。實體完整性主要是用來規範關聯表主鍵的使用原則。參考完整性是用來規範外來鍵的使用原則[1]。

實體關聯模型使用實體與關聯圖形符號元件描述實體資料之間的關係，是目前資料庫分析和設計的工具。其中實體是從真實世界中辨識出來的東西，關聯是指兩個或多個實體之間所擁有的關係[1]。

用實體關聯模型繪製的資料庫設計圖稱為實體關聯圖（ERD）。常見的實體關聯模型圖形符號如表 1 所示。屬性可分為最基本的單元值屬性型態、多個屬性所構成的複合屬性型態、可填入多重值的多重值屬性型態和用來識別實體實例的鍵屬性型態。

實體關聯圖 ERD 範例如圖 2 所示，這是一個學校資料庫的 ERD，有班（Classes）、學生（Students）、老師（Instructors）、課程（Courses）、考試（Exams）、考試結果（Results）、員工（Employees）、家長（Parents）八個實體。此 ERD 轉換成實際關聯式資料庫表格綱要如圖 3 所示。

表 1. 實體關聯圖常見符號表

實體關聯圖元件種類	圖形符號
實體	
弱實體	
關聯實體	
關聯	
識別關聯型態	
屬性	
鍵屬性	
複合屬性	
多重屬性值	

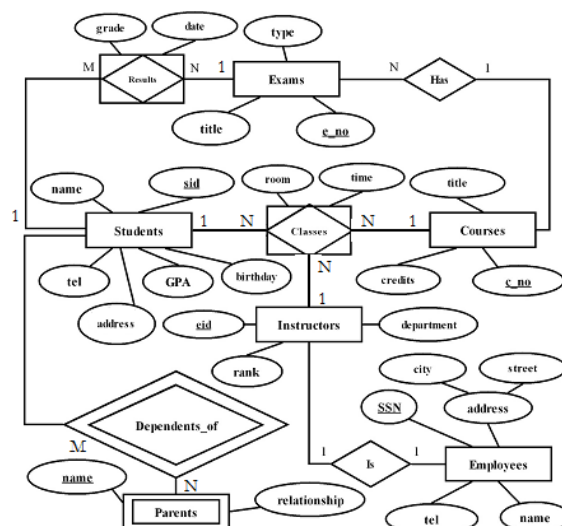


圖 2. 實體關聯圖範例

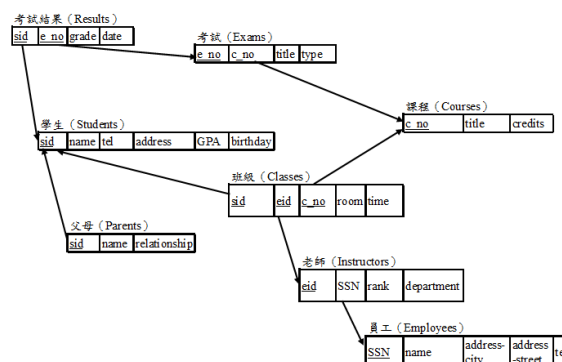


圖 3. 關聯式資料庫綱要

2.2 HBase 資料庫與資料模型

HBase 是一種開放原始碼、分散式、多版本的非關聯式資料庫[2]。其資料模型常見的名詞說明如下，其表格綱要如表 2 所示。

- (1) 列鍵是用來定義一筆資料是唯一的值，類似關聯式資料庫的 PK；
- (2) 時間戳記指定儲存格、指定版本的資料存放到區域伺服器的時間；
- (3) 欄位分成欄位家族和欄位限定符，欄位底下的所有資料以 **key-value** 的格式來儲存，其中 key 的格式為 Family:Qualifier，value 是實際儲存的資料值，以文字為主 [2][7][9]。

以表 2 為例來說，這張資料表的名稱為 Test；表格的列鍵值包括 r1 和 r2；時間戳記 ts 包括 t1、t2...等是依時間點由小到大來排序；欄位家族有 c1 和 c2，底下的所有資料依據 key-value 的方式來儲存。以 c1:q1=value1 為例，c1:q1 是 key，c1 代表欄位家族名稱，q1 代表欄位限定符名稱，value1 代表 value，是實際儲存的資料[2]。

表 2. HBase 表格 Test

Row Key	ts	Column family c1	Column family c2
r1	t1	c1:q1=value1	
	t2	c1:q2=value2	
	t3	c1:q3=value3	
	t4		c2:q1=value4
	t5		c2:q2=value5
r2	t6	c1:q1=value6	
	t7		c1:q2=value7

3. 研究方法

以下說明從現有的關聯式資料庫的資料表轉換到 HBase 資料表的作法。依據使用者的預期存取模式來設計 HBase 表格的列鍵[9]，如果無法做決定時就以新增時所產生的最少重複資料來選擇列鍵；將多個資料表合併成單一資料表時，以減少查詢時的合併時間成本。

3.1 一對一關係之範例

一位教授有一個研究室，一個研究室專屬於一個教授，兩實體之間存在一對一的關係。教授的關聯表綱要為 Professors(PID, Name, RoomID)，其中 PID 是教授的代號，Name 是教授的姓名，RoomID 是教授的研究室代號；研究室的關聯表綱要為 Rooms(RoomID, Building, Size)，其中 RoomID 是教授的研究室代號，Building 是研究室所在的位置，Size 是研究室的坪數大小。其資料建置如下：

Professors		
<u>PID</u>	Name	RoomID
P001	Dora	T2-901
P002	Fay	T2-902
P003	George	E-603

Rooms		
<u>RoomID</u>	Building	Size
T2-901	T2	2.5
T2-902	T2	2.5
E-603	E	2.6

將上述資料轉換為 HBase 表格稱為 Professors-Rooms，如表 3 所示，欄位 ts 代表 HBase 的時間戳記欄位。假設使用者比較常去查詢哪個教授在哪個研究室，所以將列鍵設為教授的代號，只要搜尋到教授的姓名就可以知道是否可以找到需要的資料。另外根據原來關聯式資料表的兩個資料表名稱，分別建立兩個欄位家族來存放資料，第一個欄位家族 Professors 包括 Name 和 RoomID 兩個欄位限定符，第二個欄位家族 Rooms 包括 Building 和 Size 兩個欄位限定符，這些都是依據關聯

式資料表的欄位名稱所作的決定。

表 3. HBase 表格 Professors-Rooms

Row Key	ts	Column family Professors	Column family Rooms
P001	t1	Professors:Name = "Dora"	
	t2	Professors:RoomID = "T2-901"	
	t3		Rooms:Building = "T2"
	t4		Rooms:Size = 2.5
P002	t5	Professors:Name = "Fay"	
	t6	Professors:RoomID = "T2-902"	
	t7		Rooms:Building = "T2"
	t8		Rooms:Size = 2.5
P003	t9	Professors:Name = "George"	
	t10	Professors:RoomID = "E-603"	
	t11		Rooms:Building = "E"
	t12		Rooms:Size = 2.6

3.2 一對多關係之範例

一位教授可指導多位研究生，一位研究生只有一位專屬指導教授，兩實體之間存在一對多的關係。教授的關聯表綱要為 Professors(PID, Name, SID)，其中 PID 是教授的代號，Name 是教授的姓名，Field 是教授的研究領域；學生的關聯表綱要為 Students(SID, Name, PID)，其中 SID 是學生的學號，Name 是學生的姓名，PID 是學生的指導教授代號。其資料建置如下：

Professors		
<u>PID</u>	Name	Field
P001	Dora	Indexing
P002	Fay	Image capture

Students		
<u>SID</u>	Name	PID
S001	Hans	P001
S002	Iris	P002
S003	Linda	P002

將上述資料轉換為 HBase 表格稱為 Students-Professors，如表 4 所示，因為一般比較常去查詢某學生是哪個教授所指導、研究哪個領域，所以將學生學號設定為 Row Key。另外根據原來關聯式資料表的兩個資料表名稱，分別建立兩個欄位家族來存放資料，第一個欄位家族 Students 包括 Name 和 Field 兩個

欄位限定符，第二個欄位家族 Professors 包括 PID 和 Name 兩個欄位限定符。

表 4. HBase 表格 Students-Professors

Row Key	ts	Column family Students	Column family Professors
S001	t1	Students:Name = "Hans"	
	t2	Students:PID = "P001"	
			Professors:Name = "Dora"
	t4		Professors:Field = "Indexing"
S002	t6	Students:Name = "Iris"	
	t7	Students: PID = "P002"	
	t8		Professors:Name = "Fay"
	t9		Professors:Field = "Image capture"
S003	t11	Students:Name = "Linda"	
	t12	Students: PID = "P002"	
	t13		Professors:Name="Fay"
	t14		Professors:Field = "Image capture"

3.3 多對多關係之範例

學生可以選多門課，課程可被很多學生選，所以兩實體之間存在多對多的關係，此關係必須轉換為一關聯實體表格才能實做出來。學生的關聯表綱要為 Students(SID, Name, Dept)，SID 代表學號，Name 代表學生姓名，Dept 代表學生所屬科系。課程的關聯表綱要為 Courses(c_no, Title, Credits)，c_no 代表課程代號，Title 代表課程名稱，Credits 代表課程學分數。關聯實體表格選課表的關聯表綱要為 Classes(SID, c_no)，SID 代表學號，c_no 代表課程代號。

Students

<u>SID</u>	Name	Dept
S001	Hans	IM
S002	Iris	IM
S003	Linda	IM

Courses

<u>c_no</u>	Title	Credits
C001	Information Hiding	3
C002	MMDB	3

Classes

<u>SID</u>	<u>c_no</u>
S001	C001
S002	C001
S002	C002
S003	C002

上述資料轉換為 HBase 表格 Classes，如表 5 所示。因為有時候會以學生的學號來查詢學生的資訊，有時候也會以課程的代號來查詢課程的資訊，所以將列鍵設為「學號+課號」。根據原來兩個關聯式資料表名稱 Students 和 Courses，分別建立兩個欄位家族來存放資料，第一個欄位家族 Students 包括 Name 和 Dept 兩個欄位限定符，第二個欄位家族 Courses 包括 Title 和 Credits 兩個欄位限定符。

表 5. HBase 表格 Classes

Row Key	ts	Column family Students	Column family Courses
S001 C001	t1	Students:Name="Hans"	
	t2	Students:Dept="IM"	
	t3		Courses:Title="Information Hiding"
	t4		Courses:Credits=3
S002 C001	t5	Students:Name = "Iris"	
	t6	Students:Dept = "IM"	
	t7		Courses:Title="Information Hiding"
	t8		Courses:Credits=3
S002 C002	t9	Students:Name = "Iris"	
	t10	Students:Dept = "IM"	
	t11		Courses:Title="MMDB"
	t12		Courses:Credits=3
S003 C002	t13	Students:Name="Linda"	
	t14	Students:Dept = "IM"	
	t15		Courses:Title="MMDB"
	t16		Courses:Credits=3

4. 結論

本文提出一套轉換原則與方法，把傳統的關聯式資料庫資料表轉換成 HBase 的非關聯式相關資料表。可以將實際運作的實務關聯式資料庫裡的大量資料轉換到 HBase 去做只需要查詢快速但不求資料精確的應用，例如空氣 PM 2.5 污染濃度對身體傷害的程度或等級的查詢。

參考文獻

- [1] 陳會安，**新觀念資料庫系統—理論與設計實務**，學貫行銷股份有限公司，2013。
- [2] 陸嘉恒，**Hadoop 實戰技術手冊**，佳魁資訊，2014。
- [3] Apache Hadoop
https://zh.wikipedia.org/wiki/Apache_Hadoop
- [4] Apache HBase
https://zh.wikipedia.org/wiki/Apache_HBase
- [5] Apache Hive
https://zh.wikipedia.org/wiki/Apache_Hive
- [6] Chang F., Dean J., Ghemawat S., Hsieh W., Wallach D., Burrows M., Chandra T., Fikes A., Gruber R., "Bigtable: A Distributed Storage System for Structured Data," *Journal of ACM Transactions on Computer Systems (TOCS)*, Volume 26, Issue 2, Article No. 4, June 2008.
- [7] Lars George 著，辜天元等譯，**HBase 技術手冊**，基峰資訊股份有限公司，2012。
- [8] MapReduce
<https://zh.wikipedia.org/wiki/MapReduce>
- [9] Nick Dimiduk、Amandeep Khurana 著，謝磊譯，**HBase: 搞定 Big Data: NoSQL 實戰**，博碩文化，2014。
- [10] Zhang N., Zheng G., Chen H., Chen J., Chen X., "HBaseSpatial: a Scalable Spatial Data Storage Based on HBase," *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 644-651, 2014.