Lesson 00 Part 01 資料庫與大數據整合

SQL Server R 服務系列之導讀說明

從大數據與深度學習的角度來看，R 與 Python 是種非常適合分析數據的電腦語言，微軟自 SQL Server 2016 開始，將 R 語言列入資料庫中的標準服務後，新版 SQL Server 2017 更加入 Python 為標準機器學習語言，從以下的安裝畫面就可以看到 Database Engine Services 中包含 Machine Learning Services(In-Database) 的 [R] 與 [Python] 的選項，合併在資料庫安裝過程之中。另外微軟也允許使用者，下載與安裝獨立的版本自行安裝。

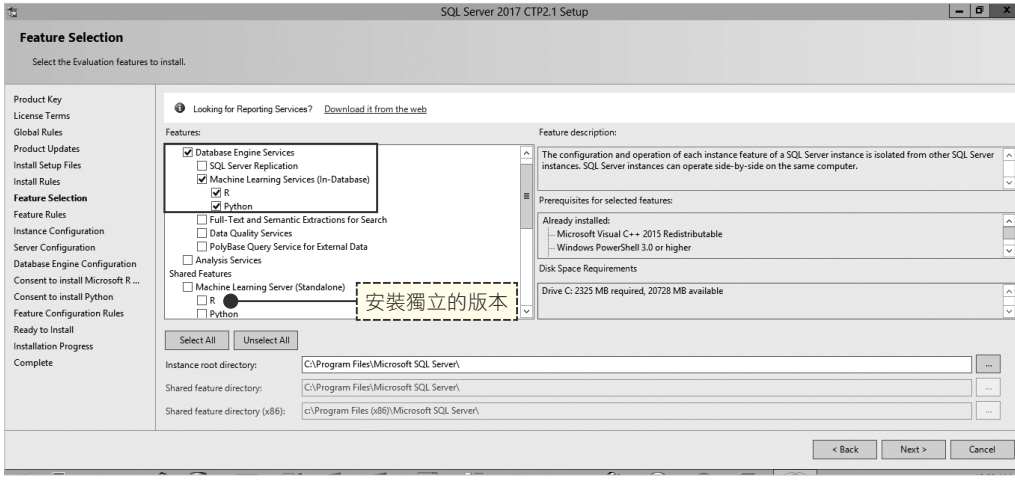


圖 1 安裝 SQL Server 過程中檢視的機器學習語言選項

R 語言之所以被許多資料科學家使用，主要原因就是內建豐富的統計與數據分析圖形功能，加上也是一種開源軟體，支援自行開發與社群分享的套件 (Packages)，可以讓 R 語言的功能，擴增到許多面向。例如預測消費者行為、股市分析與機器學習模型建立等等，都可以藉由 R 語言輕鬆達成。

Lesson 01 Part 01 資料庫與大數據整合

安裝 SQL Server R 服務

針對許多 SQL Server 用戶來說，整合 R 語言已經不用再像以往一樣，僅可以透過 RStudio 搭配 ODBC 去連接 SQL Server，進行資料抓取，然後再進行分析。現在有更棒的選擇，就是從 SQL Server 2016 開始，就可以從光碟安裝整合 SQL Server 引擎的 In-Database R 語言伺服器，然後在 SQL Server 環境中整合 R 語言分析數據。

► 案例說明

若是要安裝 R 語言，直接整合在 SQL Server 的引擎，請選擇安裝 SQL Server 2016/2017 安裝檔案的 [New SQL Server stand-alone installation or add feature to an existing installation] 版本，就可以看到安裝的步驟，此版本是由 Microsoft 公司提供整合 SQL Server 引擎的 R 語言版本。

過程中請小心若是選擇安裝檔案為，New Machine Learning Server(Standalone) installation，所安裝的 R 版本是獨立於 SQL Server 引擎，就不易整合 SQL Server 數據與 T-SQL 直接進行分析。

► 實戰解說

由於要使用較新版本的 R，建議下載 SQL Server 2017 的安裝光碟，它可以從 https://www.microsoft.com/en-us/sql-server/sql-server-2017 網址取得，本案例將使用最新的 R 版本進行內容說明。

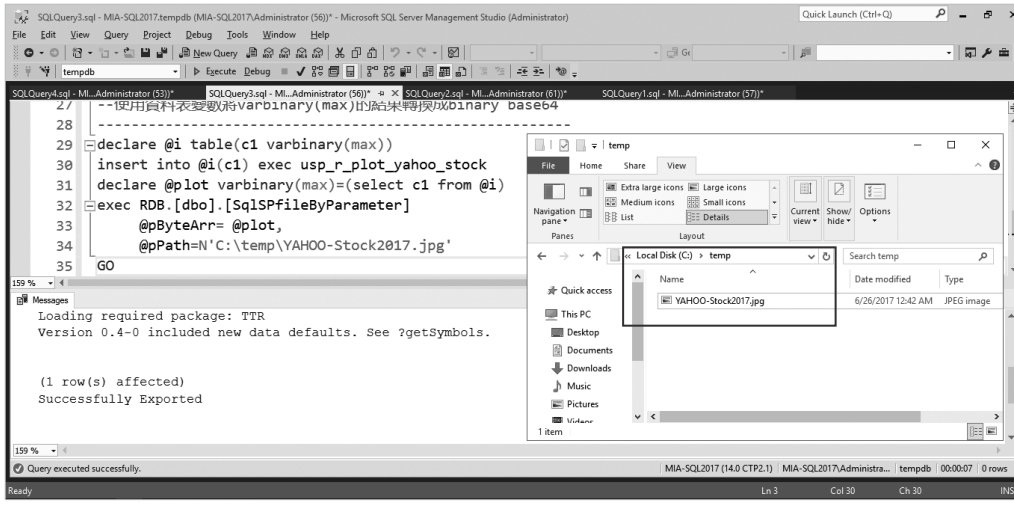


圖 7 成功執行 SQLCLR 搭配 SQL Server R 將圖片匯出到作業系統

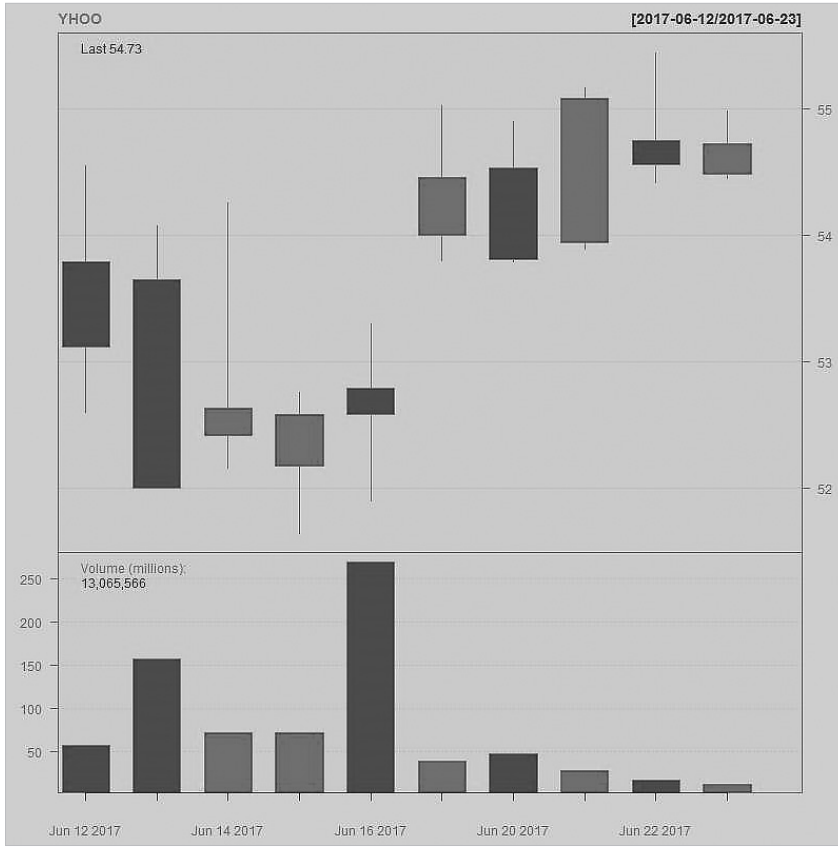


圖 8 檢視輸出的結果

Lesson 05 Part 01 資料庫與大數據整合

報表服務與 SQL Server R 呈現微軟過去一年股價資料圖

任何的資料庫新功能，最重要的部分就是終端呈現，微軟 SQL Server 的 SSRS(SQL Server Reporting Services)，就是最經濟與快速的呈現選擇。新版的 SSRS 具有許多功能，如整合 Power BI 與 Mobile Report。本範例將介紹如何整合之前的兩大範例包含有 [使用 SQL Server R 服務之 sp_execute_external_script 劃出股價圖] 與 [活用 SQL Server R 服務在 4 秒內將過去十年微軟股價載入到資料庫]，讓 SSRS 輕鬆將 SQL Server R 服務所抓取的資料，呈現出來。

► 案例說明

要從 SQL Server 的 SSRS(SQL Server Reporting Services) 繪製 SQL Server R 所抓取的資料，過程中在報表服務僅需要執行兩段 T-SQL，就可以分別 SQL Server R 匯出明細資料與 T-SQL 將 R 產生的 varbinary 資料，繪製到 SSRS 的 Image 元件。這樣的整合可以是將以下的元件應用，發揮到極致的層級。

- SQL Server R
- SQL Server T-SQL
- SQL Server Reporting Service

► 實戰解說

首先準備 SQL Server R 抓取微軟股價過去一年資料程式，該程式主要是使用 quantmod 套件搭配 getSymbols 方法從 Google API 取得微軟股票資訊，以下的程式中有一個重點就是需要將日期轉換成輸出資料行，預設的 R 服務抓取的股價資料日期是沒有辦法當成資料行。首先會取出 dfstock 的 row name 再將該 row name(就日期)，使用 cbind(column bind) 的方式，整合原來的 columns。

07 完美整合 SQL Server R 與 Database Mail 遞送數據與圖表

當使用 T-SQL 的 sp_execute_external_script 預存程序，去整合 SQL Server 與 R 服務時，最重要的環節就是直接從 SQL Server 資料庫中取出數據後，馬上轉給 R 服務，進行分析與預測，這樣可以省下大量數據在 SQL Server 與 R 服務之間傳遞的時間。本範例將使用 sp_execute_external_script 直接整合 T-SQL 查詢結果，將結果轉給 R 服務進行分析，最後藉由 SQL Server 的 database Mail 的功能，傳送給指定的收件者。

► 案例說明

在這個案例，會用到 SQL Server R、SQL CLR、SQL Server Database Mail、SQL Server T-SQL 等技術，其中 SQLCLR 主要是 SQL Server R 繪製的圖片匯出到指定路徑，再交給後續的 Database Mail 進行傳送。

首先，要先了解怎樣使用 sp_execute_external_script 預存程序抓取由 SQL Server T-SQL 傳遞過來的資料集，然後再將取得的結果，轉給 SQL Server R 的繪圖功能，繪製出整體數據的走向，最後使用 SQLCLR 將結果轉換成作業系統的檔案圖片，並且使用 SQL Server Database Mail 夾檔案的方式送出。

► 實戰解說

開始這樣的一連串技術整合之前，先來了解每個環節的運作方式，首先，是如何使用 sp_execute_external_script 預存程序，抓取由 SQL Server T-SQL 傳遞過來的資料集。這一段是很重要的部分，可以試試看以下的範例，可以將現在時間、資料庫名稱與版本傳遞給 SQL Server R 當成 data frame，再使用簡易 R 的 print 輸出結果。

最後就可以從 C:\temp\usp_r_plot_decision_tree.jpg 看到使用決策樹建立的預測模型，該模型中可以看出來影響最深的因素，就是是否有相同電信業，意思就是說國內網內互打免費的方式，該比較強的因素會去左右使用者，是否願意繼續使用該電信業者的服務（SameTelecom）。其次重要的因素就是每月撥打的分鐘數（ARPU），再其次就是年收入（YearIncome），最後就是性別（Gender）。

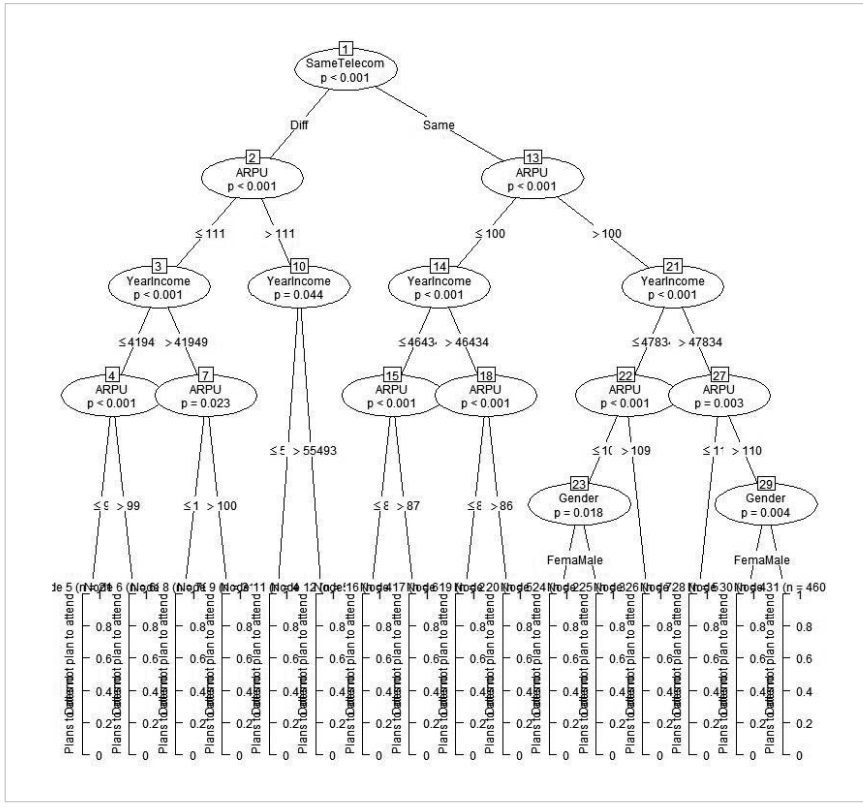


圖 3 檢視決策樹模型

► 注意事項

在執行上述程式的過程中，如果看到以下的錯誤，就是表示沒有安裝 party 套件，過程中需要使用 R.exe 安裝指定的 party 套件。

```
Msg 39004, Level 16, State 20, Line 20
A 'R' script error occurred during execution of 'sp_execute_external_script' with
HRESULT 0x80004004.
```

10 使用 SQL Server R 作為網路爬蟲抓取台灣銀行與國際匯率資料

技術可貴在於精進與提升，在筆者數年前撰寫的《SQL Server2005 資料庫程式開發達人手冊第二版》一書當中，曾經使用到 XML 技術搭配 SQLCLR 實作，取得 [立陶宛] 地區的匯率交換。過程中使用以下的方式，拿到匯率交換的資訊。

◈ T-SQL 搭配

- 預存程序
- SQLCLR
- Web Services(ASMX)
- http://webservices.lb.lt/ExchangeRates/ExchangeRates.asmx

這可說是當時最簡易的網路資料擷取方式，因為整合 SQLCLR 直接在 SQL Server 2005 的引擎中，處理資料庫的資訊下載與解析的自動化作業。

► 案例說明

隨著時間來到 SQL Server 整合 R 語言之後，就可以更簡易的方式，解決上述相同的問題，以下就是整個處理的簡易過程。其中最大的改變就是 SQLCLR 更換成 SQL Server R 技術，然後藉由 R 的豐富套件，快速完成之前的工作。

◈ T-SQL

- sp_execute_external_script
- SQL Server R 加上套件 (httr) 與 (XML)
- Web Services(ASMX)
- http://webservices.lb.lt/ExchangeRates/ExchangeRates.asmx

13 活用 SQL Server R 語言整合作業系統 WMIC 來監控硬碟空間

R 語言整合 SQL Server 資料庫，除了可以分析資料庫中的數據，也可以活用它的 system 函數去擷取所在系統中的資訊進行分析與處理，以下就是使用 R system 函數去呼叫 Windows 作業系統 DIR 指令去找出有多少 SQL 開頭的目錄與檔案名稱，過程中使用 intern=T 指定 R 將結果回傳，並將結果儲存到指定的變數，該變數經過檢查不是 dataframe。

```
df<-system("cmd.exe /c dir  c:\\SQL* ", intern = T)
print(df);
is.data.frame(df);
```

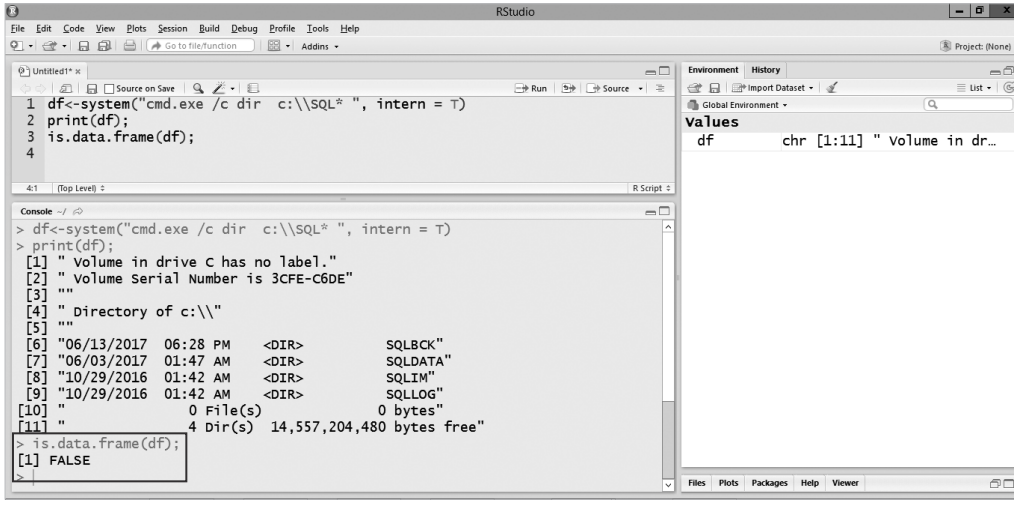


圖 1 使用 R 語言的 system 函數呼叫 Windows 作業系統中的 DIR 指令

若是需要將作業系統執行的結果，轉換成 R 語言的 dataframe 來整合微軟的 SQL Server 預存程序 sp_execute_external_script，則需要使用 R 的 pipe() 函數進行轉換，以下就是使用 R pipe() 函數將結果轉換成 dataframe 方式，方便後續資料儲存。

Lesson 01

Part 02 資料庫開發技術聖殿

NULL 處理技巧之不同 NOT IN NOT EXISTS EXCEPT 使用方式比較

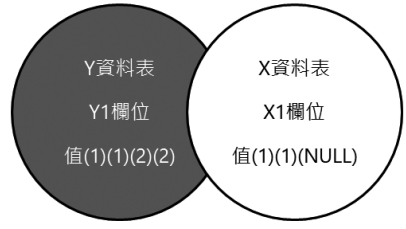
當需要在兩個資料表中找出差異值時，許多人第一直覺就是使用 [NOT IN]，通常這樣做不會有甚麼問題，但若碰到 [NOT IN] 的子查詢資料值有 NULL 時，就全盤皆輸，意思就是找不出任何差異。這樣在小量資料可以藉由眼力觀察的狀況下，還可以找出這樣寫法 [NOT IN] 的危險地方，但是碰到背景程式，或是資料量多時，幾乎無法察覺到這樣的危險。所以告訴自己不要再用 [NOT IN] 去找出兩邊資料差異。

► 案例說明

下列的案例有兩個資料表，一個是 [X]，一個是 [Y]。其中 [X] 資料表包含三筆資料，其中一筆是 NULL，而所謂的 NULL 就是沒有值，它跟空白、零或是空字串都不一樣。當使用子查詢時，要檢視那些 [Y] 資料表的值不存在於 [X] 資料表，大部分的人都會用以下的方式，就是 [NOT IN]。

兩資料表比較時候處理NULL問題

- 左邊資料表為Y，該資料表Y1欄位
  - 希望找出Y.Y1欄位值不存在於右邊X.X1欄位




```
-- 希望要找出答案是 2,2，但答案卻不如預期
 SELECT y1 FROM Y
 WHERE  y1 NOT IN
(SELECT x1 FROM X);
```

圖 1 找出不存在於另一資料表欄位的值

08 不為人知的 OPENQUERY 秘密功能

當需要藉由 SQL Server 執行個體去查詢其他資料庫，如 Oracle 或 MYSQL 時，許多人會用 Linked Server 的方式，連接到異質的資料庫，進行 SELECT 陳述式執行，但鮮少有人知道 OPENQUERY 搭配 Linked Server 函數，基本上是可以允許從 SQL Server 端，藉由 Linked Server 與 OPENQUERY 的方式，直接進行遠端資料的異動，包括 INSERT/UPDATE 與 DELETE。

以下的例子，就是使用 SQL Server 結合 Linked Server 與 OPENQUERY 的方式，直接異動 ORACLE 資料庫中的 HR.EMPLOYEES 與 MYSQL 的資料。

► 案例說明

◈ 異動 ORACLE 資料庫

要實作這個範例，首先，需要在該 x64 位元 SQL Server 的機器中安裝 Oracle 驅動程式，該 Linked Server 的驅動程式 ODAC121024Xcopy_x64.zip 可以從以下 URL 取得：

http://www.oracle.com/technetwork/database/windows/downloads/index-090165.html

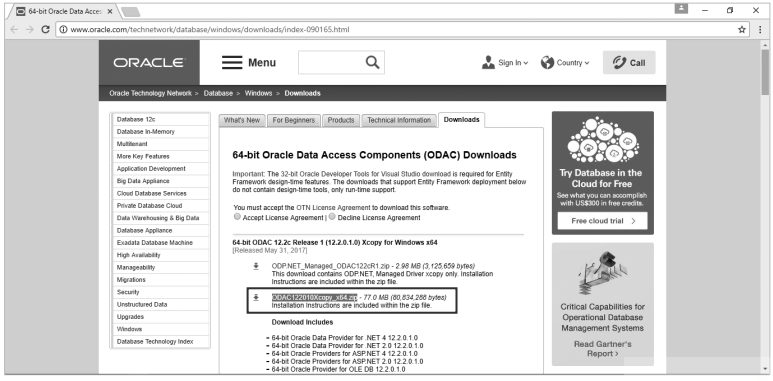


圖 1 取得 ODAC122010Xcopy_x64 程式

Lesson 11 使用 BCP 程式匯出資料庫影像資料，無須撰寫 ADO.NET 或是 JDBC 程式

大家都知道，T-SQL 的 OPENROWSET 函數，可以快速載入二進位元資料到作業系統，如相片、影片。有人會詢問是否有機會，不要撰寫 ADO.NET 或是 JDBC 程式，就直接使用 SQL Server 內建工具，將資料庫中的影像、影片等 BLOB 資料匯出，並且根據原始的檔案名稱，自動儲存到作業系統的目錄？

基本上，這問題著實不簡單，但是可以使用 BCP 工具來完成，對大家來說，使用 BCP 多半是用在文字、UNICODE 或是 NATIVE 的轉換資料表內容，極少用在 BLOB 的影像、影片資料，以下就是 BCP 的使用說明。

```
--BCP 使用說明
c:\temp>bcp/?
usage: bcp {dbtable | query} {in | out | queryout | format} datafile
  [-m maxerrors]            [-f formatfile]          [-e errfile]
  [-F firstrow]             [-L lastrow]             [-b batchsize]
  [-n native type]          [-c character type]      [-w wide character type]
  [-N keep non-text native] [-V file format version] [-q quoted identifier]
  [-C code page specifier]  [-t field terminator]    [-r row terminator]
  [-i inputfile]            [-o outfile]             [-a packetsize]
  [-S server name]          [-U username]            [-P password]
  [-T trusted connection]   [-v version]             [-R regional enable]
  [-k keep null values]     [-E keep identity values]
  [-h "load hints"]         [-x generate xml format file]
  [-d database name]        [-K application intent]  [-l login timeout]
```

► 案例說明

首先，來練習使用 BCP 工具，如何將資料庫文字資料匯出。過程中搭配 T-SQL 輸出資料庫中的查詢結果，到 C:\temp\emp.txt 並且使用文字格式進行匯出。

15 SQL Server 2016 之 STRING_SPLIT 快速解決斷行斷字需求

當看到 SQL Server 2016 有一個 STRING_SPLIT 函數時，迫不急待應用在自己的現有系統上面，竟然超出想像的快速，對比傳統方式，成本竟然為 99:1。以往 SQL Server 要斷行斷字時，需要使用 CHARINDEX 與 SUBSTRING，自己撰寫使用者自訂函數如下，該函數需要使用迴圈的方式，才可以針對指定字元進行字串的斷行斷字。

```
-- 使用傳統方式斷行斷字

CREATE FUNCTION [dbo].[fnSplitString]
(
    @string NVARCHAR(MAX),
    @delimiter CHAR(1)
)
RETURNS @output TABLE(id int identity,splitdata NVARCHAR(MAX)
)
BEGIN
    DECLARE @start INT, @end INT
    SELECT  @start = 1, @end = CHARINDEX(@delimiter, @string)
    WHILE   @start < LEN(@string) + 1 BEGIN
        IF  @end = 0
            SET @end = LEN(@string) + 1

        INSERT INTO @output (splitdata)
        VALUES(SUBSTRING(@string, @start, @end - @start))
        SET @start = @end + 1
        SET @end = CHARINDEX(@delimiter, @string, @start)

         delete from @output where splitdata=''

    END
```