Externe R-Skripte ausführen mit SQL Server und ML Services

Kathegorie
SQL Server

Aaron
Priesterroth

Das **R Project** ist eine Entwicklungsumgebung für statistische und graphische Komputation. Sie steht auf einer Vielzahl von Plattformen zur Verfügung und ist in den vergangenen Jahren zu einem der wichtigsten Werkzeuge in diesem Gebiet geworden. Dies beruht vor allem auf der hohen Modularität, mit deren Hilfe Benutzer sog. Packets verwalten können. Mit diesen Packets können der Umgebung (angepasst an den entsprechenden Arbeitsauftrag) spezielle Funktionalitäten zugewiesen und diese mit allen anderen Benutzern geteilt werden. So stehen jedem Benutzer nach groben Schätzungen zum momentanen Stand (Mai 2020) ca. 17.500 Pakete mit individueller Funktionalität zur Verfügung.
Auch im Umfeld des SQL Server findet R einen hohen Anwendungsbedarf: statistische Analyse, Aufbereitung und Verarbeitung von Daten sowie die graphische Darstellung sind vor allem (aber nicht nur) im Zusammenhang mit den Machine Learning Services häufig im Einsatz.
Die allgemeine Benutzung eines R-Skripts, genauer die Ausführung von R-Skripten mit Ein- und Ausgabe-Parametern, möchten wir in diesem Artikel genauer unter die Lupe nehmen.

Voraussetzungen

Um externe R-Skripte mit Hilfe des SQL Server ausführen zu können, müssen folgende Voraussetzungen erfüllt sein:

- Es muss eine SQL Server Instanz mit den **SQL Server Machine Learning Services** zur Verfügung stehen. Auf dieser Instanz muss die Sprache **R** installiert sein.
- Die Berechtigung zum Ausführen von externen Skripten muss gegeben sein. Der Vorgang wird hier in einem Artikel von Microsoft beschrieben.
- Es muss sichergestellt sein, dass der **SQL Server Launchpad Service** gestartet ist.

Mainzer Datenfabrik GmbH
Rotterdamer Platz 2, 55120 Mainz
info@madafa.de
madafa.de
+49 6131 3331612

Ausführen eins R-Skripts

Einer der wichtigsten Aspekte in der Benutzung von R in Zusammenhang mit dem SQL Server ist das Ausführen eines Skripts. Dies geschieht mit Hilfe der gespeicherten Prozedur sp_execute_external_script. Die Prozedur

- initialisiert die Laufzeit-Umgebung von R im Kontext des SQL Servers,
- stellt die Daten für die Ausführung bereit,
- kümmert sich um die sichere Verwaltung von Benutzer-Sessions und
- liefert die Ergebnisse der Ausführung an den Benutzer zurück.

Für die Darstellung der Ausführung möchten wir folgendes simples R-Skript betrachten:

```
print("Hello World!", quote = FALSE)
```

Bei Ausführung gibt das Skript die Zeichenkette **”Hello World!“** auf der Konsole aus.

Um das oben beschriebene Skript auszuführen, müssen Sie die folgenden Schritte befolgen:

- **Öffnen Sie SQL Server Management Studio** (oder ein Alternativ-Programm) und verbinden Sie es mit einer Instanz. Diese Instanz muss die oben genannten Voraussetzungen erfüllen.
- Mit Hilfe des Knopfes "**Neue Abfrage**“ können Sie ein neues Abfrage-Fenster öffnen.
- Innerhalb des Abfrage-Fensters kann nun mit Hilfe des folgenden Befehls das beispielhafte R-Skript ausgeführt werden:

```
EXECUTE sp_execute_external_script
    @language = N'R',
    @script = N'
print("Hello World!", quote = FALSE)
'
```

- Nach der Ausführung erscheint die Ausgabe im **Messages**-Fenster:

```
STDOUT message(s) from external script:
Hello World!
```

Ausführen eines R-Skripts mit Ein- und Ausgabe

Die Prozedur **sp_execute_external_script** erwartet eine optionale Datenmenge als Eingabe. Diese Eingabe wird in den meisten Fällen in der Form einer validen t-SQL Abfrage realisiert.

Um die Benutzung von Input-Daten zu verdeutlichen, betrachten wir die folgende Tabelle:

```
CREATE TABLE ProductInventory (
  product_name nvarchar(100) NOT NULL,
  amount_available int NOT NULL,
  amount_sold int NOT NULL);

INSERT INTO ProductInventory([product_name], [amount_available], [amount_sold])
VALUES
    ('Banana', 73, 112),
    ('Apple', 17, 201),
    ('Orange', 3, 461),
    ('Pear', 89, 183);
```

Die Tabelle hält drei Spalten, den Namen eines Produkts, die verfügbare Kapazität und die Anzahl der abgeschlossenen Verkäufe (TotalSales). Sie muss auf der benutzten Instanz erstellt und befüllt werden.

Die als Eingabe gewählte Datenmenge wird bei Aufruf der Prozedur vom SQL Server automatisch in der R Umgebung zur Verfügung gestellt. Nach der Verarbeitung durch das R-Skript wird eine Ergebnismenge als Dataframe (das Äquivalent einer Tabelle in R) zurück geliefert. Um diese implizite Konvertierung genauer zu verstehen, betrachten wir im Folgenden ein R-Skript, dessen Eingabe gleichzeitig die Ausgabe ist:

```
EXECUTE sp_execute_external_script
    @language = N'R',
    @script = N'OutputDataSet <- InputDataSet;',
    @input_data_1 = N'SELECT SUM([amount_sold]) FROM
ProductInventory;'
WITH RESULT SETS(([TotalSales] INT NOT NULL));
```

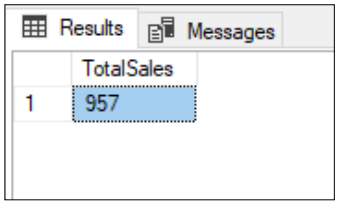

Mainzer Datenfabrik GmbH
Rotterdamer Platz 2, 55120 Mainz
info@madafa.de
madafa.de
+49 6131 3331612

Wobei

- @language die verwendete Sprache bezeichnet (in diesem Fall R)
- @scriptdas verwendete Skript bezeichnet (in diesem Fall die simple Zuweisung der Eingabe als Ausgabe). Hierbei sind vor allem die Namen der standardmäßigen Aus- und Eingabewerte **OutputDataSet** und **InputDataSet** interessant.
- @input_data_1 die Datenmenge, die als Eingabe verwendet wird, bezeichnet (in diesem Fall die Abfrage auf den **ProductInventory** Table).
- WITH RESULT SETS(...) die implizite Konvertierung, der von R gelieferten Dataframes in das "neue" SQL Server ResultSet, beschreibt.

Nach der Ausführung des Skripts wird folgende Ausgabe zurückgegeben: die Anzahl der verkauften Produkte in Summe (TotalSales).

Namen des In- und OutputDataSet-Parameters ändern



In manchen Situationen kann es sehr hilfreich sein, den Namen des In- und OutputDataSets zu ändern. Diese werden standardmäßig mit **InputDataSet** und **OutputDataSet** bezeichnet. Dabei ist besonders wichtig, die richtige Groß- und Kleinschreibung zu beachten, da es sich bei R um eine **case-sensitive** Sprache handelt.

Um den Namen der Ein- bzw. Ausgabe-Parameter zu verändern, kann bei der Ausführung der Prozedur sp_execute_external_script der Parameter @input_data_1_name bzw. @ output_data_1_name verwendet werden.

Am Beispiel des zuvor ausgeführten Skripts sehen die Änderungen wie folgt aus:

```
EXECUTE sp_execute_external_script
    @language = N'R',
    @script = N'my_out <- my_in;',
    @input_data_1 = N'SELECT 123 AS [ColumnName];'
    @input_data_1_name = N'my_in',
    @output_data_1_name = N'my_out'
WITH RESULT SETS(([NewColumnName] INT NOT NULL));
```

Die Parameter der Ein- und Ausgabe werden nun mit den Bezeichnern my_in und my_out identifiziert.

Weiterführende Schritte

Das Ausführen eines externen R-Skripts mit Hilfe des SQL Servers ist ein von Microsoft sehr einfach und zugänglich gestalteter Vorgang. Die draus resultierenden Möglichkeiten für den Einsatz der Sprache in der Server-Umgebung sind nahezu endlos und können den Arbeitsablauf jedes Anwenders, gerade im Bereich der Analyse, enorm erleichtern.

Weiterführend sind vor allem der Umgang mit Objekten und Datenstrukturen in R von Vorteil. Weitere Informationen dazu können Sie direkt auf der Website des R Projects nachlesen.

ausführen, execute, execute r script, machine learning, ml
services, r, r project, r script, r skript, sql server

Mainzer Datenfabrik GmbH
Rotterdamer Platz 2, 55120 Mainz
info@madafa.de
madafa.de
+49 6131 3331612