extreme Datamining mit Oracle R Enterprise

Matthias Fuchs
ISE Information Systems Engineering GmbH
Gräfenberg

Schlüsselworte
ISE, Oracle R Enterprise, Engineered System, Exadata, Datamining, CRAN, Statistik, SQL, Storage Server, Database Machine, Exalytics

Einleitung
Es ist seit einiger Zeit möglich statistische R Berechnungen innerhalb der Oracle Datenbank durchzuführen. Durch die R Berechnung innerhalb der Datenbank verändern sich die Arbeitsabläufe und die Performance steigt unter bestimmten Voraussetzung beträchtlich. Auf großen Datenmengen ergeben sich dadurch neue Möglichkeiten. Es wird in dem Vortrag gezeigt wie man Oracle R Enterprise auf einem Oracle Engineered System (Exadata) installiert und konfiguriert. Danach werden die Unterschiede bei der Verwendung von standard R scripten und Oracle R embedded scripten dargestellt. Am Ende wird auf Optimierungsmöglichkeiten bei der Ausführung in der Datenbank eingegangen.
In Zusammenarbeit mit eoda, ein auf dataming spezialisiertes Unternehmen, führt die ISE GmbH Projekte in Umgebungen mit Oracle Datenbanken durch. Dabei kommen oft Exadata und Exayltics Systeme zum Einsatz.

„R“

R ist eine freie Programmiersprache für statistisches Rechnen und statistische Grafiken. R gilt zunehmend als die statistische Standardsprache, sowohl im kommerziellen als auch im wissenschaftlichen Bereich. Durch den modularen Aufbau und die große Vielfalt von Erweiterungen (Paketen) bietet die Sprache viele Einsatzmöglichkeiten in der Statistik. Ob lineare oder nichtlineare Modellierung, Zeitreihenanalyse oder Clusteranalyse mit „R“ können fast alle Analysen durchgeführt werden.

„R“ und Big Data

Immer kürzere Produktlebenszyklen, der Trend zur Individualisierung sowie die fortschreitende Digitalisierung nahezu aller Geschäftsbereiche erhöhen die Menge der vorhandenen Daten und gleichzeitig die Notwendigkeit, intelligent mit dem Rohstoff Daten umzugehen. Die zu analysierenden Daten sind meist strukturiert in einer Datenbank abgelegt. Erreichen die Datenmengen mehrer Terrabyte, man kann von Big Data sprechen, kommen oft Oracle Datenbanken zum Einsatz.

Eine Kombination aus Oracle Datenbank und „R“ zur Analyse von strukturierten Daten ist daher eine Schlussfolgerung. Genau dieser Ansatz soll im Folgenden beschrieben werden.

Die Grundlage: Datamining in der Oracle Datenbank

Die Analyse von Daten umfasst mehrere Schritte. Die meiste Zeit geht vor der eigentlichen Analyse bei der Datenaufbereitung verloren. Es sind Exporte und Konvertierungen der Rohdaten durchzuführen. Diese werden dann wiederum auf weitere Systeme kopiert, um mit seperaten Analysewerkzeugen direkten Zugriff zu haben. Während dieses Ablaufes geht viel Zeit verloren.

Zusätzlich sind weitere Hardwareresourcen erforderlich. Diese Schritte entfallen, wenn die Daten an Ort und Stelle, in der Datenbank, verarbeitet werden. Zusätzlich greifen vorhanden Security und Compliance Richtlinien in der Datenbank und müssen nicht auf anderen Systemen repliziert werden.

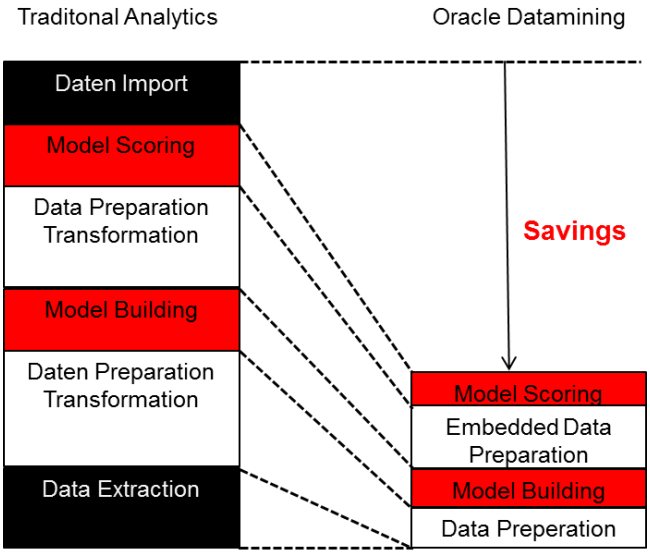

Abb. 1: Optimierungen beim in Oracle „R“ Database Datamining

Die erzielten Ergebnisse liegen ebenfalls wieder in der Datenbank und müssen nicht aufwendig importiert werden.

Zusätzlich zum einfacheren Datenhandling kommen Performancesteigerungen beim Analysieren der Daten. Je nach verwendetem Algorithmus ist mit einer deutlicher Beschleunigung beim Model Scoring oder Model Building zu rechnen. Die Verwendung von etablierten Standards für die Rechteverwaltung und Zugriffssteuerung, brauchen ebenfalls nicht verändert werden bzw. sind bereits vorhanden.

Die Erweiterung: Datamining mit Oracle „R“ Enterprise

Oracle hat die Verwendung von „R“ innerhalb der Datenbank transparent implementiert. Dies ist sowohl bei einer Installation auf Standard Hardware, als auch bei sogenannten Engineered Systems, wie die Oracle Exadata Database Machine, möglich.

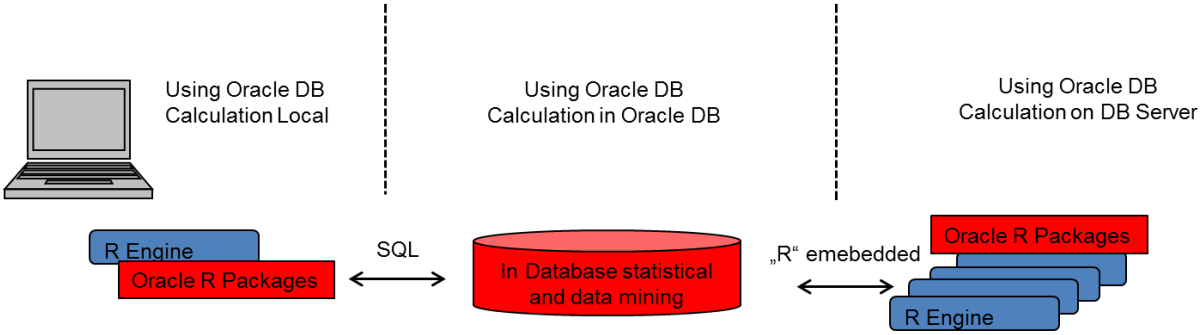

Abb. 1: Oracle „R“ Szenarios

„R“ Prozesse können auf drei Arten mit einer Oracle DB verarbeitet werden:

Die Berechnungen laufen auf einen unabhängigen Server bzw. Client und nur die Daten werden direkt aus der Oracle Datenbank geladen. Eine aufwendige Konvertierung der Daten in z.B. XML Datenstrukturen oder CSV Files entfällt. Es können alle R CRAN Packete verwendet werden.

Alternativ kann man die Berechnungen auch direkt auf dem Datenbankserver starten. Dies erfolgt z.B. aus PL/SQL Prozeduren heraus. Der Vorteil besteht darin das keinerlei Netzwerkverkehr entsteht. Nur die Ergebnisse werden zum Client übertragen. Es können auch hier alle R Packete verwendet werden.

Als letzte Möglichkeit Analysen mit R durchzuführen gibt es von Oracle speziell angepasste R Prozeduren. Dabei wurden die Pakete für die Ausführung in der Datenbank optimiert. Dadurch ergeben sich deutliche Performancesteigerungen gegenüber der Verwendung der „normalen“ R Pakete.

Der Performance Gewinn ist bei Berechnungen in der Datenbank mit den angepassten R Packeten am höchsten. Aber auch beim Zugriff auf R in der Datenbank werden deutliche Geschwindigkeitssteigerungen erreicht. Bei allen Arten der Verwendung können immer alle R Packete angesprochen werden, die im Comprehensive R Archive Network (CRAN) veröffentlicht sind und mit der verwendeten R Basis Version lauffähig sind. Natürlich können auch eigene, selbst entwickelte R Pakete eingespielt werden.

Die Optimierung: „R“ on Oracle Exadata

Die Exadata Database Machine war das erste Engineered System bei Oracle. Die Oracle Exadata Database Maschine ist die optimale Lösung für Data Warehouse Anwendungen, als auch für OLTP Systeme. Die Kombination aus speziellen Oracle Storage Server Software und der bekannten Oracle Datenbank, alles basierenden auf Standard Hardware von Sun, ermöglicht extreme Geschwindigkeiten in ein hochverfügbaren und hoch sichern Umgebung. Geliefert, als ein optimiertes und vorkonfiguriertes Gesamtsystem aus Software, Rechnern und Storage ist das System einfach und schnell zu implementieren und verwendbar für alle Arten von Applikationen die in einem Geschäftsumfeld eingesetzt werden.

Bei der Verwendung einer Exadata als Datenbank Server. ergeben sich weitere Vorteile für Statische Analysen. Der hohe Datendurchsatz kann beim Lesen der Daten effektiv genutzt werden. Dadurch ergeben sich teilweise stark verkürzte Analysezeiten. Das Parallelisieren von „R“ Prozessen kann, basierend auf der Oracle Real Application Cluster Datenbanktechnologie, problemlos auf mehreren Nodes stattfinden. Einer Analyse von großen Datenmengen im Terrabyte Bereich steht somit nichts mehr im Wege.

Die „R“ Installation wird von Oracle angepasst und für die Exadata Database Machine zertifiziert und optimiert. Ein Einsatz im geschäftlichen Umfeld ist somit problemlos möglich.

Erfahrungen und Beispiele

Im Rahmen des Vortrages wird geschildert, was bei der Erstinstallation von R in einer Datenbank zu beachten ist. Beispiele werden gezeigt, wie „R“ code in einer Datenbank ausgeführt werden kann und was zu beachten ist um die Leistung einer Oracle Exadata auszunutzen.

Es werden R Codes lokal im z.B. R Studio gestartet, auf der Exadata ausgeführt und die Ergebnisse zurück an den Client gesendet. Dieses Script wird dann angepasst um es innerhalb einer PLSQL Prozedur laufen zu lassen.

Kontaktadressen:

Matthias Fuchs
ISE Information Systems Engineering GmbH
Gewerbepark Hüll 4
D-91322 Gräfenberg

Telefon: +49 (0) 172-8288751
Fax: +49 (0) 9192-9929-22
E-Mail matthias.fuchs@ise-informatik.de
Internet: www.ise-informatik.de

Herbert Rossgoderer
ISE Information Systems Engineering GmbH
Gewerbepark Hüll 4
D-91322 Gräfenberg

Telefon: +49 (0) 172- 8244478
Fax: +49 (0) 9192-9929-22
E-Mail herbert.rossgoderer@ise-informatik.de
Internet: www.ise-informatik.de

Oliver Bracht
eoda Heiko Miertzsch & Oliver Bracht GbR
Ludwig-Erhard-Straße 10
D-34131 Kassel

Telefon: +49 (0) 561/202724-40
Fax: +49 (0) 561/202724-30
E-Mail: info@eoda.de
Internet: http://www.eoda.de