R Statistik im Oracle Produktstack

Matthias Fuchs
ISE Information Systems Engineering GmbH
Nürnberg

Schlüsselworte
ISE, Oracle R Enterprise, ORE, Exadata, Datamining, CRAN, Statistik, Exalytics, OBIEE

Einleitung
Oracle hat die OpenSource Programmiersprache R in mehreren Produkten integriert. Oracle bietet damit eine Möglichkeit, Analytic und Visualisierungen im Datenbank und Big Data Umfeld durchzuführen. Im wird dargetellt, wie die Integration durchgeführt wurde. Zusätzlich werden Hinweise gegeben, wann der Einsatz sinnvoll ist.

„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 mehrere 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.

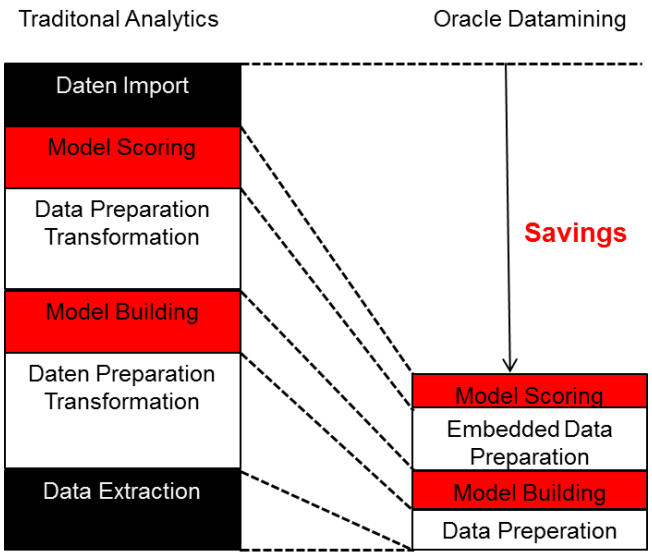


Abb. 1: Optimierungen beim in Oracle „R“ Database Datamining

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.

ORE - R in der Datenbank

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

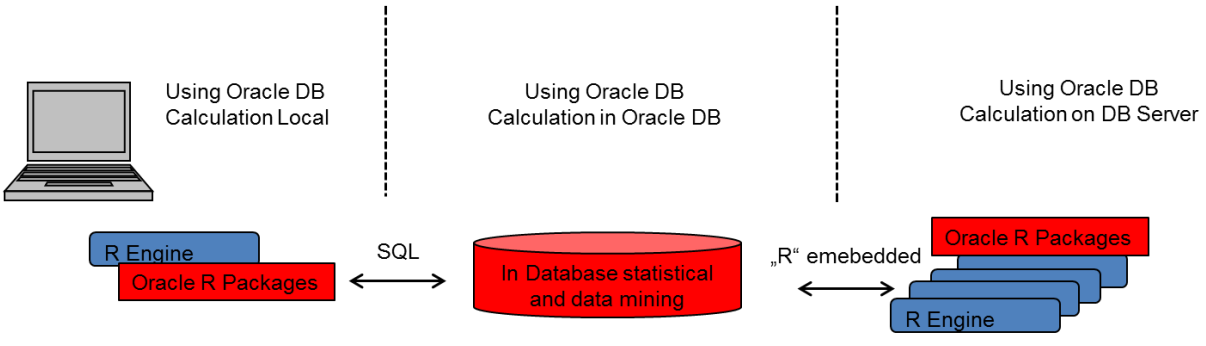


Abb. 1: Oracle „R“ Szenarios

„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 Pakete verwendet werden.

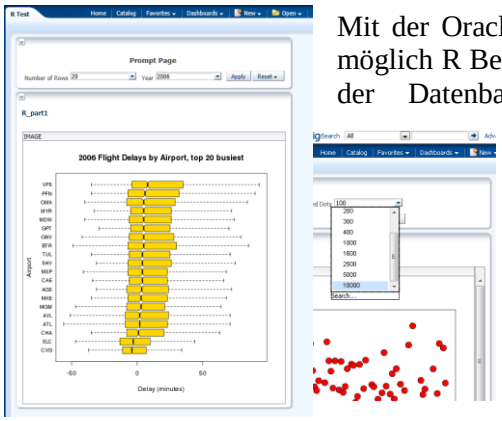
Als letzte Möglichkeit Analysen mit R durchzuführen gibt es von Oracle speziell angepasste R Prozeduren. Dabei wurden die Pakete für die Ausführung in der Datenbank optimiert. Dadurch ergeben sich deutliche Performancesteigerungen gegenüber der Verwendung der „normalen“ R Pakete.

ORE – Parallelität, Memory und Performance

Die Vorzüge der Integration liegen darin, dass die Performance steigt. Dies kann natürlich nur funktionieren, wenn Funktionen der Datenbank genutzt werden. Dies kann über R Scripten oder PLSQL Scripten erfolgen.

Die Funktionen werden erläutert und mit Beispielen Optimierungsponential dargestellt. Ebenfalls ist es auf der einen Seite sinnvoller R Funktionen zu nutzen, auf der anderen Seite auf Funktionen der Datenbank zurückzugreifen. Hier muss ein Verständnis für das Gesamtsystem aufgebaut werden.

„R“ und Oracle BI



Mit der Oracle Business Intelligence Enterprise Edition (OBIEE) ist es möglich R Berechnungen direkt auf den Analyse Daten oder über SQL in der Datenbank auszuführen. Ebenso können alle Arten von Visualisierungen aus „R“ direkt in einem Dashboard dargestellt werden. Somit werden die Möglichkeiten der Analyse deutlich erhöht.

„R“ und Hadoop - ORAAH

Oracle R Advanced Analytics For Hadoop (ORAAH) ist eine weitere Entwicklungen im Rahmen von R von Oracle. Mit ORAAH können MAP/Reduce Jobs gestartet werden. Ebenso kann HIVE oder HDFS angesprochen werden. Ein Datanaustausch istzwichen Datenbank und Hadoop möglich.

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