Datenanalysen auf Enterprise-Niveau mit Oracle R Enterprise

Dr. Nadine Schöne, Dr. Michael Haupt, Negib Marhoul
Oracle Deutschland BV & Co. KG
Potsdam

Schlüsselworte
Oracle R Enterprise, Datenanalysen, Data Analysis, Reporting, Data Mining, R Programming Language, Statistik, Statistics, Graphics, Graphische Darstellung

Einleitung
Stellen Sie sich vor, Sie bemerken am Ende der DOAG, dass Sie etwas verloren haben, z.B. Ihre Uhr. Die Uhr könnte in sich in einem der 19 Vortragsräume oder irgendwo auf dem Konferenzgelände befinden, im Hotel, oder in einem Taxi. Auf dem Konferenzgelände können Sie sofort persönlich suchen, und die Suche mit netter Hilfe sogar parallelisieren. Im Hotel könnten Sie anrufen, und anhand der Taxiquittungen lässt sich auch bestimmt herausfinden, welchen Taxifahrer Sie kontaktieren müssen. Mit etwas Glück finden Sie Ihre Uhr – und wahrscheinlich auch noch ein paar Dinge, nach denen Sie gar nicht gesucht haben.
Die Analyse großer Datenmengen sieht ähnlich aus: Die zu analysierenden Daten sind meist auf verschiedene Datenbanken und Filesysteme verteilt. Wenn Analysetools direkt auf Daten zugreifen können (so wie Sie selbst das Konferenzgelände durchsuchen können), beschleunigt dies die Analyse. Auch Parallelisierung ist ein Schlüssel zu größerer Geschwindigkeit. Und genau wie bei der Suche nach Ihrer Uhr können Sie mit Mining-Verfahren auch Muster finden, nach denen Sie gar nicht explizit gesucht hatten.

Datenanalysen im Enterprise
Statistik und Mining-Verfahren sind in der Regel iterativ: Daten werden gesammelt, identifiziert und aufbereitet, um dann analysiert zu werden. Ausgehend von den Analyseergebnissen werden erneut Daten gesammelt, und die nächste Runde beginnt (s. Abb. 1).

Für die Implementierung solcher Datenanalysen, die über ein Standard-Reporting hinausgehen, bietet sich Oracle R Enterprise (ORE) an. ORE ist speziell auf Data Mining und die Analyse großer Datenmengen ausgerichtet.

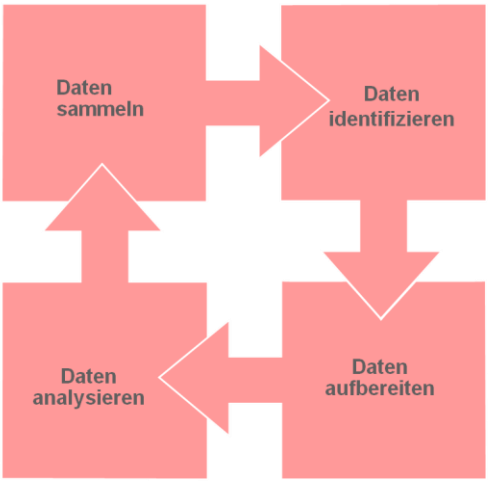


Abb. 1: Iteratives Vorgehen bei Statistik und Mining-Verfahren

R und Oracle R Enterprise (ORE)

ORE ist eine von Oracle entwickelte Variante der Open-Source-Programmiersprache R. Open-Source R wurde von Statistikern speziell für die Datenanalyse entwickelt. R ist eine statistische Workbench, ein Data-Science-Ökosystem, und DIE *lingua franca* für Data Science.

Inzwischen wird Open-Source R weltweit insbesondere in der Forschung, aber auch immer mehr in anderen Bereichen, z. B. der Analyse von Kundendaten, verwendet. Dies beruht nicht zuletzt auf deren vielfältigen Möglichkeiten der Visualisierung. Die Funktionalität von R wird permanent durch die Nutzer selbst erweitert: Gekapselte Funktionalität kann in Form von R-Packages für alle Nutzer auf einem zentralen Server zum Download bereitgestellt werden.

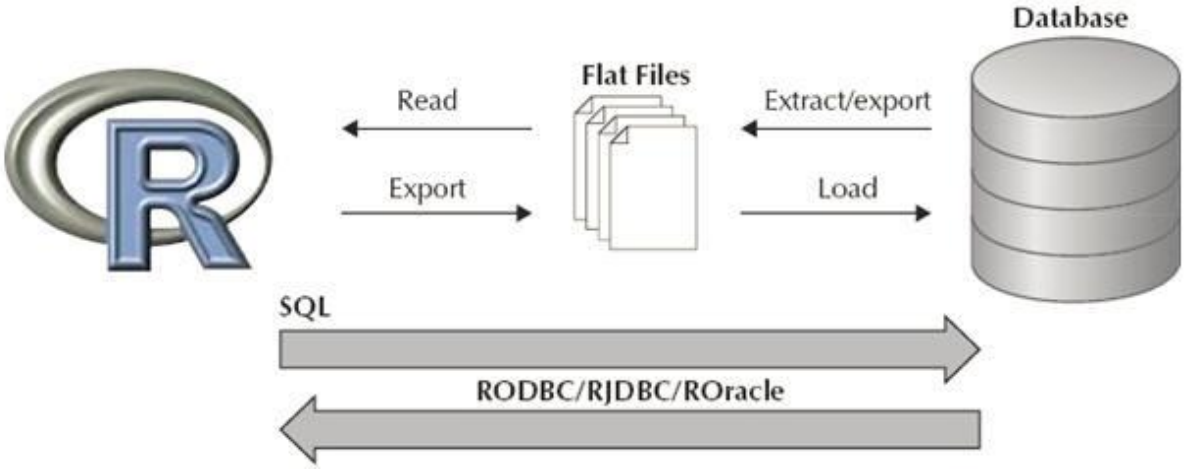


Abb. 2: Aspekte herkömmlicher R/Datenbank-Interaktionen

Open-Source R wird in der Regel auf einem Client (PC oder Notebook) installiert. Daten müssen zur Verarbeitung mittels Flat File Export erst von der Datenbank in den R-Workspace importiert werden (s. Abb. 2). Bei größeren Datenmengen vermindert dies die Analysegeschwindigkeit erheblich. Weiterhin ist die Parallelisierung von Berechnungen mit R nur händisch möglich, automatische Parallelisierung ist nicht vorgesehen.

Im Gegensatz zu R nutzt ORE nicht nur eine Client-Engine, sondern nutzt durch eine zweite R-Engine (u.U. sogar mehrere) auf dem Datenbankserver auch die Rechenkraft der Datenbank. Somit entfällt mit ORE der Datenexport, wodurch ORE ungleich performanter als R ist. Außerdem ist eine Parallelisierung der Anfragen möglich, was die Performance weiter steigern kann.
Moment – Datenbank? Braucht man da nicht auch noch SQL? Die ORE Client Engine wandelt in ihrem Transparency Layer Anfragen an die Datenbank in SQL um, so dass die explizite Übergabe von SQL-Befehlen entfällt. Um mit ORE Datenanalysen zu implementieren braucht man kein SQL; Kenntnisse in R reichen aus.
Mit ORE lässt sich sämtlicher nativer R Code verwenden. Auch alle über CRAN downloadbaren R-Pakete sind nutzbar. Genau wie R bietet auch ORE eine große Vielfalt an graphischen Darstellungen, die einfach und schnell erstellt werden können.

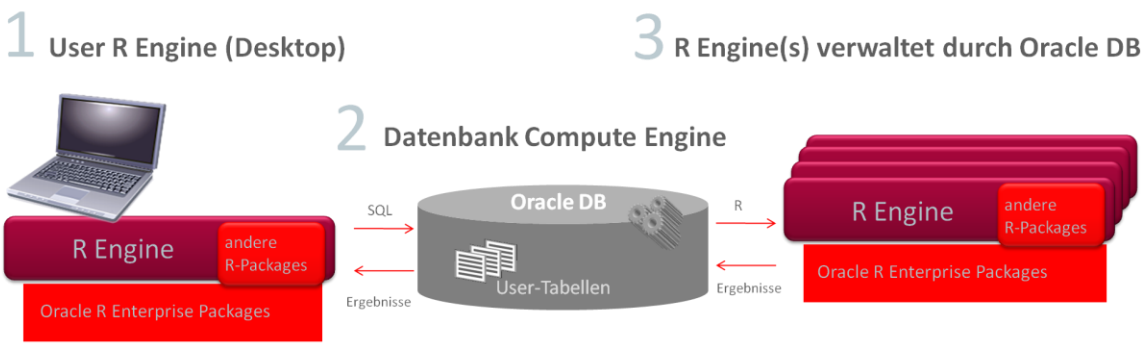


Abb. 2: "Collaborative Execution"-Modell mit ORE

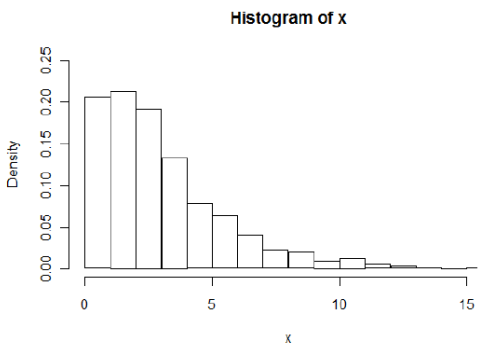
Wie einfach und schnell eine Grafik erzeugt werden kann, zeigt das folgende Beispiel.

```
set.seed(25) ; x <- rchisq(1000, df = 3)
```

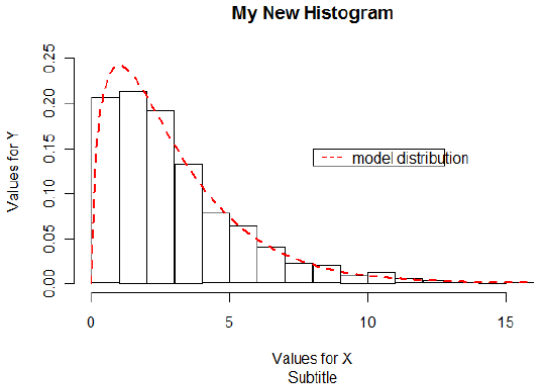
In den Vektor x werden 1000 Zufallszahlen nach der Chi-Quadrat-Verteilung mit einem Freiheitsgrad von 3 abgelegt. Danach soll ein Histogramm der Daten visualisiert werden. Mit dem Befehl

```
hist (x, freq=FALSE, ylim = c(0, 0.25))
```

wird das Histogramm mit der Dichte von x und der Verteilung zwischen 0 und 0,25 erzeugt. Das zweite Beispiel ist eine Erweiterung des ersten Histogramms und zeigt wie Graphiken mit weiteren Informationen wie Legenden, Beschriftungen und weiteren Grafiken angereichert werden können.




```
hist(x, freq = FALSE, ylim = c(0, 0.25))
```




```
hist(x, freq = FALSE, ylim = c(0, 0.25),
     main="My New Histogram", sub ="Subtitle",
     xlab="Values for X",
     ylab="Values for Y")
curve(dchisq(x, df = 3), col = 2, lty = 2,
     lwd = 2, add = TRUE)
legend(8,0.15,"model distribution",
       col="red", lty=2)
```

Oracle Labs und FastR

Die Open-Source-R-Implementierung hat die Eigenschaft, dass sie R-Code Schritt für Schritt ausführt („interpretiert“), was hinsichtlich der Ausführungsgeschwindigkeit nicht optimal ist. R-Entwickler wenden aus diesem Grund häufig die Strategie an, die Performance-kritischen Teile ihrer Anwendungen in C, C++ oder Fortran neu zu schreiben und von R aus aufzurufen.

Oracle Labs ist die Forschungsabteilung von Oracle. Hier wird auf allen Ebenen des Oracle-Stacks an Innovationen gearbeitet, die schließlich per Technologietransfer in Produktgruppen überführt werden. Die Virtual Machine Research Group bei Oracle Labs befasst sich insbesondere mit der effizienten Implementierung von Programmiersprachen. Eines der Projekte ist FastR, eine Neuimplementierung von R in Java.

FastR baut auf Truffle und Graal auf—gleichfalls Projekte bei Oracle Labs, die zum Ziel haben, die Implementierung von Programmiersprachen zu vereinfachen und gleichzeitig sehr gute Ausführungsgeschwindigkeit zu erreichen. Anders als Open-Source-R übersetzt FastR häufig ausgeführte Teile des R-Codes in Maschinencode für die Hardware, so dass der Aufwand der Interpretierung wegfällt.

Truffle, Graal und FastR sind Open-Source-Projekte.

Weitere Informationen

Neugierig geworden? Dann klicken Sie sich durch untenstehende Links oder rufen Sie uns an.

ORE Discussion Forum:
https://community.oracle.com/community/developer/english/business_intelligence/data_warehousing/r

Oracle Advanced Analytics:
http://www.oracle.com/technetwork/database/options/advanced-analytics/index.html

ORE-Blog:
https://blogs.oracle.com/R/

FastR:
https://bitbucket.org/allR/fastR

Graal/Truffle:
https://wiki.openjdk.java.net/display/Graal/Main

Oracle Labs im OTN:
http://www.oracle.com/technetwork/oracle-labs/index.html

Kontaktadresse:

Dr. Nadine Schöne
Telefon: +49 (0) 331-200 7190
E-Mail nadine.schoene@oracle.com

Dr. Michael Haupt
Telefon: +49 (0) 331-200 7277
E-Mail michael.haupt@oracle.com

Negib Marhoul
Telefon: +49 (0) 331-200 7217
E-Mail negib.marhoul@oracle.com

Oracle Deutschland BV & Co. KG
Schiffbauergasse 14
D-14467 Potsdam
Fax: +49 (0) 12-345 6788
Internet: www.oracle.com