

Holger Schwarz

Anfragegenerierende Systeme

VIEWEG+TEUBNER RESEARCH

Holger Schwarz

Anfragegenerierende Systeme

Anwendungsanalyse, Implementierungs-
und Optimierungskonzepte

VIEWEG+TEUBNER RESEARCH

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<<http://dnb.d-nb.de>> abrufbar.

Habilitationsschrift Universität Stuttgart, 2009

D 93

1. Auflage 2010

Alle Rechte vorbehalten

© Vieweg+Teubner Verlag | Springer Fachmedien Wiesbaden GmbH 2010

Lektorat: Ute Wrasmann | Britta Göhrisch-Radmacher

Vieweg+Teubner Verlag ist eine Marke von Springer Fachmedien.

Springer Fachmedien ist Teil der Fachverlagsgruppe Springer Science+Business Media.

www.viewegteubner.de



Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg

Druck und buchbinderische Verarbeitung: STRAUSS GMBH, Mörlenbach

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Printed in Germany

ISBN 978-3-8348-1298-8

Vorwort

Das vorliegende Buch basiert auf den Ergebnissen von Forschungsarbeiten am Institut für Parallele und Verteilte Systeme der Universität Stuttgart, an denen ich in den vergangenen Jahren beteiligt war. In der ersten Phase dieser Arbeiten lag das Hauptaugenmerk auf der Integration heterogener Daten in einem Data Warehouse und den darauf aufbauenden Analysemöglichkeiten. In meiner Dissertation zum Thema „Integration von Data Mining und Online Analytical Processing: Eine Analyse von Datenschemata, Systemarchitekturen und Optimierungsstrategien“ habe ich zentrale Aspekte einer integrierten Betrachtungsweise von Data-Mining- und OLAP-Analysen bearbeitet. Hierbei spielten auch die für diese Varianten der Datenanalyse relevanten Datenzugriffe auf den Data-Warehouse-Datenbestand sowie Ansätze zu deren Optimierung eine wichtige Rolle.

Die zweite Phase der Forschungsarbeiten war einerseits geprägt durch eine Ausweitung der Anwendungspalette (für die Datenzugriffe genauer zu analysieren waren) und andererseits durch eine Weiterentwicklung der betrachteten Optimierungsstrategien. Die Palette berücksichtigter Anwendungsbereiche reichte von der Datenanalyse auf Basis eines Data Warehouse bis hin zu Workflowsystemen und Anwendungen für die semantische Suche. Im dem vom Bundesministerium für Bildung und Forschung geförderten Projekt nova-net wurden beispielsweise Techniken zur Generierung von Webanwendungen sowie Techniken zur themenspezifischen Suche von Informationen im Web, zum webbasierten Erfassen von Informationen und der zugehörigen Informationsaufbereitung entwickelt. In enger Kooperation mit IBM wurden die Datenzugriffe in datenintensiven Workflows analysiert und Ansätze zu deren Optimierung entwickelt. Verwandte Optimierungsfragestellungen wurden darüber hinaus beispielsweise in dem von der DFG geförderten Projekt CEOPS bearbeitet.

So heterogen sich die Anwendungsszenarien in den erwähnten und weiteren Projekten auch darstellten, so hat sich doch ein gemeinsames Muster gezeigt: In allen Anwendungen werden Anweisungen für Datenzugriffe generiert. Diese Beobachtung lieferte den Anlass, zu einer intensiveren Beschäftigung mit dem Thema der Anfragegenerierung und hat schlussendlich zu diesem Buch geführt. Wichtige Fragestellungen, denen ich dabei nachgegangen bin, sind Gemeinsamkeiten und Unterschiede der verfolgten Generierungsansätze, die Einbindung der

Anfragegenerierung in die jeweilige Systemarchitektur sowie die Defizite der generierten Anfragen und vielversprechende Optimierungsansätze.

Das vorliegende Buch beruht auf meiner im Jahr 2009 an der Fakultät für Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart eingereichten Habilitationsschrift, deren Entstehen ohne eine fortwährende intensive wissenschaftliche Förderung und Kooperation nicht möglich gewesen wäre. An erster Stelle gebührt Herrn Prof. Dr. Bernhard Mitschang Dank für seine stete Unterstützung meiner Forschungsarbeiten und vielfältige Denkanstöße zu den behandelten Themen. Herrn Professor Johann-Christoph Freytag danke ich für wichtige Anmerkungen zu einer ersten Fassung meiner Habilitationsschrift sowie für seine Bereitschaft, das Korreferat im Zuge meines Habilitationsverfahrens zu übernehmen. Ebenso großen Dank schulde ich den vielen Kollegen aus den verschiedenen Forschungsvorhaben, die durch Diskussionen und Implementierungsarbeiten einen unverzichtbaren Beitrag zu meiner Arbeit geleistet haben. Insbesondere sind hier Herr Dr. Severin Beucker, Herr Dr. Klaus Fichter, Herr Mihály Jakob, Herr Fabian Kaiser, Herr Dr. Dierk-Oliver Kiehne, Herr Dr. Tobias Kraft, Herr Dr. Albert Maier, Herr Dr. Christoph Mangold, Herr Oliver Suhre und Herr Marko Vrhovnik zu nennen.

Der mit einer intensiven Forschungsarbeit verbundene Zeitaufwand hinterlässt auch im persönlichen Bereich seine Spuren. Meiner Familie danke ich für das entgegengebrachte Verständnis. Insbesondere gilt mein Dank meiner Frau Sabine sowie meinen Töchtern Julia und Daniela für ihre Geduld und ihre moralische Unterstützung.

Stuttgart, im April 2010

Holger Schwarz

Inhaltsverzeichnis

| | | |
|----------|--|-----------|
| 1 | Einleitung | 1 |
| 2 | Grundlagen | 5 |
| 2.1 | Zentrale Anwendungsklassen | 5 |
| 2.1.1 | Datenbankanwendungen | 5 |
| 2.1.2 | Information-Retrieval- und Webanwendungen | 7 |
| 2.2 | Datenbearbeitungsanweisungen | 9 |
| 2.2.1 | Sprachen für Datenbearbeitungsanweisungen | 9 |
| 2.2.2 | Einbettung von Datenbearbeitungsanweisungen | 11 |
| 2.2.3 | Komplexität von Datenbearbeitungsanweisungen | 12 |
| 2.3 | Entwicklungsprozesse für Anwendungsprogramme | 13 |
| 2.3.1 | Phasen der Softwareentwicklung | 13 |
| 2.3.2 | Erstellungszeitpunkt der Datenbearbeitungsanweisungen | 14 |
| 2.4 | Systemarchitekturen | 16 |
| 2.4.1 | Datenbankanwendungen | 16 |
| 2.4.2 | Webanwendungen | 17 |
| 2.4.3 | Information-Retrieval-Anwendungen | 21 |
| 2.4.4 | Service-Orientierte Architekturen | 23 |
| 2.5 | Anfragegenerierende Systeme | 23 |
| 2.5.1 | Begriffsklärung und allgemeines Systemmodell | 24 |
| 2.5.2 | Abgrenzung zu anderen Systemklassen | 25 |
| 2.5.3 | Generierung von Datenbearbeitungsanweisungen | 26 |
| 2.6 | Zusammenfassung | 27 |
| 3 | Verwendung anfragegenerierender Systeme | 29 |
| 3.1 | Motivation für die Verwendung anfragegenerierender Systeme | 29 |
| 3.1.1 | Flexibilität in der Reaktion auf Eingaben | 30 |
| 3.1.2 | Flexibilität in der Anpassung an die Datenverarbeitungskomponente | 31 |
| 3.1.3 | Komplexitätsreduktion | 31 |
| 3.1.4 | Anpassbarkeit und Wartbarkeit | 32 |
| 3.2 | Fragestellungen anfragegenerierender Systeme im Überblick | 33 |
| 3.2.1 | Systemklassifikation | 33 |

| | | |
|----------|--|-----------|
| 3.2.2 | Generierungsansätze | 34 |
| 3.2.3 | Optimierungsansätze | 35 |
| 3.3 | Zusammenfassung | 35 |
| 4 | Szenarien und Systembeispiele | 37 |
| 4.1 | Business Intelligence | 37 |
| 4.1.1 | Systemarchitektur | 38 |
| 4.1.2 | Generierung von Datenbearbeitungsanweisungen | 40 |
| 4.2 | Datenmanagement in datenintensiven Workflows | 40 |
| 4.2.1 | Systemarchitektur | 43 |
| 4.2.2 | Generierung von Datenbearbeitungsanweisungen | 44 |
| 4.3 | Generierung eines Repositories | 45 |
| 4.3.1 | Systemarchitektur | 45 |
| 4.3.2 | Generierung von Datenbearbeitungsanweisungen | 46 |
| 4.4 | Generierung datenintensiver Webanwendungen | 47 |
| 4.4.1 | Systemarchitektur | 47 |
| 4.4.2 | Generierung von Datenbearbeitungsanweisungen | 48 |
| 4.5 | Semantische Suche | 49 |
| 4.5.1 | u38 | 49 |
| 4.5.2 | EXPOSE | 51 |
| 4.6 | Zusammenfassung | 55 |
| 5 | Klassifikation anfragegenerierender Systeme | 57 |
| 5.1 | Klassifikationskriterien | 57 |
| 5.1.1 | Generierungszeitpunkt | 57 |
| 5.1.2 | Anfragesprache | 59 |
| 5.1.3 | Zusammenhang der Anfragen | 60 |
| 5.1.4 | Variabilität der Anfragen | 61 |
| 5.1.5 | Komplexität der Anfragen | 62 |
| 5.1.6 | Klassifikationsschema | 62 |
| 5.2 | Einordnung der Systembeispiele | 63 |
| 5.2.1 | Business Intelligence | 63 |
| 5.2.2 | Datenmanagement in datenintensiven Workflows | 64 |
| 5.2.3 | Generierung eines Repositories | 65 |
| 5.2.4 | Generierung datenintensiver Webanwendungen | 65 |
| 5.2.5 | Semantische Suche | 66 |

- 5.3 Zusammenfassung 66

- 6 Ansätze zur Generierung von Datenbearbeitungsanweisungen 69**
 - 6.1 Bewertungskriterien 70
 - 6.2 Parametrisierung 72
 - 6.2.1 Vorgehensweise 72
 - 6.2.2 Bewertung 74
 - 6.3 Template-basierte Ansätze 75
 - 6.3.1 Vorgehensweise 75
 - 6.3.2 Bewertung 77
 - 6.4 Algorithmen-basierte Ansätze 78
 - 6.4.1 Vorgehensweise 78
 - 6.4.2 Bewertung 79
 - 6.5 Zusammenfassende Bewertung der Generierungsansätze 80
 - 6.6 Abgrenzung zu anderen Ansätzen der Code-Generierung 82
 - 6.7 Einordnung der Systembeispiele 84
 - 6.7.1 Business Intelligence 84
 - 6.7.2 Datenmanagement in datenintensiven Workflows 84
 - 6.7.3 Generierung eines Repositories 85
 - 6.7.4 Generierung datenintensiver Webanwendungen 86
 - 6.7.5 Semantische Suche 86
 - 6.8 Zusammenfassung 86

- 7 Ansätze zur Optimierung generierter Datenbearbeitungsanweisungen 89**
 - 7.1 Überblick 89
 - 7.2 Analyse des Optimierungsbedarfs 90
 - 7.3 Klassifikation der Optimierungsansätze 91
 - 7.4 Optimierung bei der Anwendungsentwicklung und in der Applikation 93
 - 7.4.1 Einzeloptimierung 94
 - 7.4.2 Homogene Optimierung 96
 - 7.4.3 Heterogene Optimierung 97
 - 7.5 Optimierung durch das Datenmanagementsystem 97
 - 7.5.1 Einzeloptimierung 98
 - 7.5.2 Homogene Optimierung 99

| | | |
|----------|--|------------|
| 7.5.3 | Heterogene Optimierung | 99 |
| 7.6 | Optimierung durch eine separate Komponente | 100 |
| 7.6.1 | Einzeloptimierung | 100 |
| 7.6.2 | Homogene Optimierung | 101 |
| 7.6.3 | Heterogene Optimierung | 101 |
| 7.7 | Zuordnung der Optimierungsansätze | 102 |
| 7.8 | Zusammenfassung | 106 |
| 8 | Vertiefung und Bewertung ausgewählter Optimierungsansätze | 109 |
| 8.1 | Anfrageoptimierung in Datenbanksystemen | 110 |
| 8.2 | Multi-Query-Optimierung in Datenbanksystemen | 113 |
| 8.2.1 | Einsatzgebiete und Weiterentwicklungen | 114 |
| 8.2.2 | Anwendbarkeit für anfragegenerierende Systeme | 116 |
| 8.3 | CGO-Ansatz zur Optimierung von Anfragesequenzen | 117 |
| 8.3.1 | Voraussetzungen und Ziele | 117 |
| 8.3.2 | Optimierungsansatz | 120 |
| 8.3.3 | Klassifikation der Optimierungsregeln | 123 |
| 8.3.4 | Heuristischer CGO-Ansatz | 128 |
| 8.3.5 | Kostenbasierter CGO-Ansatz | 130 |
| 8.3.6 | Effektivität und Effizienz der Optimierung | 136 |
| 8.3.7 | Zusammenfassung der Coarse-Grained-Optimierung | 142 |
| 8.4 | PGM-Optimierung des Datenmanagements in Workflows | 142 |
| 8.4.1 | Voraussetzungen und Ziele | 143 |
| 8.4.2 | Optimierungsansatz | 144 |
| 8.4.3 | Klassifikation der Optimierungsregeln | 150 |
| 8.4.4 | Anforderungen an eine interne Repräsentation | 151 |
| 8.4.5 | Das Prozessgraphmodell | 155 |
| 8.4.6 | Optimierungsregeln auf Basis von PGM | 160 |
| 8.4.7 | Eigenschaften des Prozessgraphmodells | 163 |
| 8.4.8 | Kontrollstrategie | 165 |
| 8.4.9 | Effektivität der Optimierung | 168 |
| 8.4.10 | Zusammenfassung der PGM-Optimierung | 172 |
| 8.5 | Bewertung der Optimierungsansätze | 173 |
| 8.5.1 | Bewertung hinsichtlich Effektivität und Ausgereiftheit | 173 |
| 8.5.2 | Bewertung im Kontext anfragegenerierender Systeme | 175 |
| 8.5.3 | Einsatzmöglichkeiten in den Systembeispielen | 178 |
| 8.6 | Zusammenfassung | 181 |

| | |
|---------------------------------------|------------|
| 9 Zusammenfassung und Ausblick | 183 |
| 9.1 Resümee..... | 183 |
| 9.2 Ausblick..... | 187 |
| 10 Literaturverzeichnis | 189 |

Abbildungs- und Tabellenverzeichnis

| | |
|--|-----|
| Abbildung 1: Zentrale Anwendungsklassen. | 8 |
| Abbildung 2: Wichtige Phasen eines Softwareentwicklungsprozesses | 13 |
| Abbildung 3: Systemarchitektur von Datenbankanwendungen | 17 |
| Abbildung 4: Architekturvarianten für Webanwendungen | 18 |
| Abbildung 5: Systemarchitektur von Webanwendungen | 20 |
| Abbildung 6: Crawler-Indexer-Architektur für Suchmaschinen | 21 |
| Abbildung 7: Abgrenzung anfragegenerierender Systeme. | 26 |
| Abbildung 8: Dreischichtenarchitektur von Business-Intelligence- Anwendungen. | 38 |
| Abbildung 9: Beispielworkflow mit Datenmanagementaktivitäten. | 42 |
| Abbildung 10: Systemarchitektur datenintensiver Workflows. | 44 |
| Abbildung 11: Architektur von SERUM | 46 |
| Abbildung 12: Architektur des Webanwendungs-Generators ALGen | 48 |
| Abbildung 13: Systemarchitektur von u38. | 50 |
| Abbildung 14: Expertensuche in EXPOSE | 52 |
| Abbildung 15: Systemarchitektur von EXPOSE | 54 |
| Abbildung 16: Zuordnung der Systembeispiele zu Anwendungsklassen | 55 |
| Abbildung 17: Softwareentwicklungsprozess und Generierungszeitpunkte | 59 |
| Abbildung 18: Klassifikationsschema | 62 |
| Abbildung 19: Übersicht der Generierungsansätze | 69 |
| Abbildung 20: Generierungsansatz Parametrisierung | 73 |
| Abbildung 21: Generierungsansatz template-basiert | 76 |
| Abbildung 22: Generierungsansatz algorithmen-basiert | 79 |
| Abbildung 23: Matrix der Optimierungsmöglichkeiten | 92 |
| Abbildung 24: Matrix zur Bewertung des Optimierungspotenzials. | 93 |
| Abbildung 25: Übersicht der Optimierungsmöglichkeiten | 103 |
| Abbildung 26: Optimierungsansätze | 109 |
| Abbildung 27: Beispiel einer Anfragesequenz. | 118 |
| Abbildung 28: Anweisungstrippel und Abhängigkeitsgraph. | 119 |
| Abbildung 29: Einzelanfrage zur Beispielsequenz. | 121 |
| Abbildung 30: Restrukturierung mit der WhereToGroup-Regel | 122 |
| Abbildung 31: Architektur eines CGO-Optimierers | 128 |
| Abbildung 32: Architektur eines kostenbasierten CGO-Optimierers. | 133 |
| Abbildung 33: Effektivität des heuristischen CGO-Ansatzes | 137 |
| Abbildung 34: Effizienz des CGO-Ansatzes | 139 |
| Abbildung 35: Effektivität des kostenbasierten CGO-Ansatzes | 141 |

| | |
|--|-----|
| Abbildung 36: Architektur der PGM-Optimierung | 145 |
| Abbildung 37: Beispiel-Workflow | 147 |
| Abbildung 38: Optimierter Beispiel-Workflow | 148 |
| Abbildung 39: Beispiel-Workflow nach der kombinierten Optimierung. | 149 |
| Abbildung 40: Klassifikation der Restrukturierungsregeln | 150 |
| Abbildung 41: Basiskonzepte des Prozessgraphmodells | 155 |
| Abbildung 42: Aufbau des generischen Aktivitätstyps | 156 |
| Abbildung 43: Hierarchische Beziehung der Aktivitätstypen | 157 |
| Abbildung 44: Aufbau einzelner Aktivitätstypen. | 158 |
| Abbildung 45: PGM-Repräsentation des Beispielworkflows | 159 |
| Abbildung 46: Restrukturierungsregel Tuple-to-Set | 161 |
| Abbildung 47: Abhängigkeiten zwischen Restrukturierungsregeln. | 165 |
| Abbildung 48: Kontrollstrategie für die PGM-Optimierung | 167 |
| Abbildung 49: Effektivität einzelner Restrukturierungsregeln | 169 |
| Abbildung 50: Optimierung des Beispielworkflows | 170 |
| Abbildung 51: Optimierung einer benutzerdefinierten Funktion. | 171 |
| Abbildung 52: Optimierung von benutzerdefinierter Funktion und Workflowbeschreibung | 172 |
| Abbildung 53: Zusammenhang zwischen Generierungszeitpunkt und gewinnbringenden Optimierungsansätzen | 176 |
| Abbildung 54: Bedeutung der Optimierungsansätze für die einzelnen Generierungsansätze | 177 |
| | |
| Tabelle 1: Einordnung der Systembeispiele | 63 |
| Tabelle 2: Bewertung der Generierungsansätze | 81 |