

Institut für Informationssysteme  
Fachgruppe Datenbanken

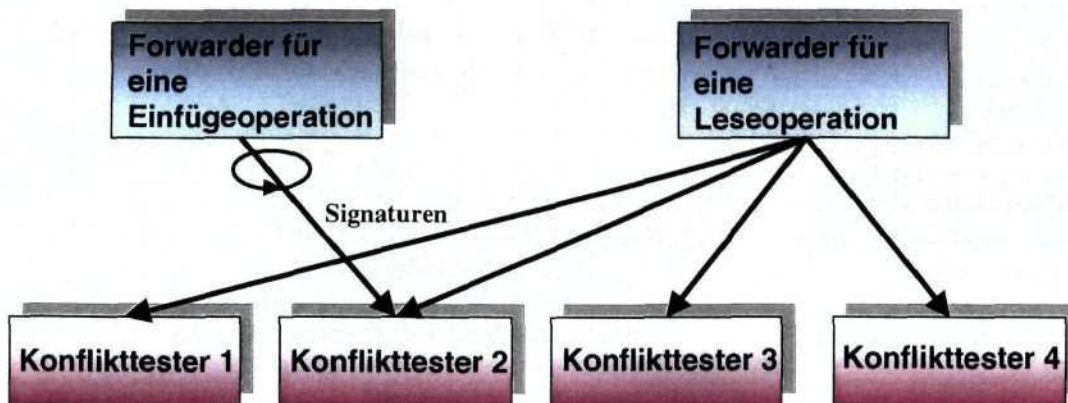
Prof. H.-J. Schek

Diplomarbeit WS99/00

---

*Verteilung der Termextraktion und des Konflikttests in der  
PowerDB Dokumentensuche*

---



Name: Martin Vogt  
Abteilung: IIIC  
Student-Nr: 94-921-814  
Email: mvogt@iiic.ethz.ch

Betreuer: Torsten Grabs  
Email: grabs@inf.ethz.ch

# Inhaltsverzeichnis

1	Einleitung .....	1
1.1	Motivation.....	1
1.2	Aufgabenstellung.....	1
1.3	Gliederung.....	2
2	Überblick über die Architekturen.....	5
2.1	Dokumentensuche.....	5
2.2	Was ist PowerDB? .....	5
2.3	Gemeinsamkeiten beider Architekturen .....	5
2.3.1	Erlaubte Operationen.....	5
2.3.2	Aufteilen der Aufträge in Subtransaktionen.....	6
2.3.2.1	Einfügeoperation.....	7
2.3.2.2	Leseoperation .....	7
2.3.3	Konflikttest.....	7
2.3.4	Verteilung der Daten auf die verschiedenen Komponenten .....	8
2.3.5	Logging.....	9
2.4	Alte Architektur .....	9
2.5	Neue Architektur.....	11
3	Dokumentensuchmaschine.....	13
3.1	Überblick über die Server und Services.....	13
3.2	DB-Schema und physisches Design .....	14
3.2.1	Tabelle ADM_PARADOCU .....	14
3.2.2	Tabelle ADM_DOCID .....	15
3.2.3	Tabelle ADM_TERMS .....	15
3.2.4	Tabelle TP_LIB .....	16
3.2.5	Tabellen IL_TP_LIB_SUBJECT, BODY, COMBINED.....	16
3.2.6	Physisches Design .....	16
3.3	Server für das Vergeben von Transaktions-Ids (GetTID) .....	17
3.4	Server für das Generieren der Signatur (GenerateSig) .....	17
3.4.1	Termextraktion: Bestimmen der Terme .....	17
3.4.2	Mapping der Terme auf die Signaturbits .....	18
3.4.2.1	Einfaches Hashing.....	18
3.4.2.2	Einfaches Hashing unter Berücksichtigung der Buchstabenposition....	19
3.4.2.3	Einfaches Hashing unter Berücksichtigung der Wortlänge .....	19
3.4.2.4	Hashing nach Pearson .....	20
3.5	Server für den verteilten Konflikttest .....	23
3.5.1	Konflikttest mit Verteilung der vollständigen Signaturen .....	23
3.5.1.1	Server Forwarder.....	24
3.5.1.2	Server DistrConfTest .....	25
3.5.1.3	Mögliche Deadlocksituation .....	27
3.5.2	Konflikttest mit Aufteilen der Signatur.....	28
3.5.2.1	Synchronisierung der Konflikttester .....	29
3.5.2.2	Server Forwarder (2).....	30
3.5.2.3	Server DistrConfTest (2).....	31
3.6	Server für das Einfügen von Dokumenten.....	34
3.6.1	Server InsertDocu .....	35
3.6.2	Server GetMaxDocId.....	36

3.6.3	Server MInsertDocu .....	37
3.6.4	Server MInsertDesc .....	37
3.6.5	Server GetTermId .....	37
3.7	Server für Leseoperation .....	38
3.7.1	Server RetrieveDocu .....	38
3.7.2	Server MRetrieveDesc .....	39
3.7.3	Server MRetrieveDocu .....	40
3.8	Plug & Play Skalierbarkeit .....	40
3.8.1	Vorbereitung .....	40
3.8.2	Informieren der RetrieveDocu-Server und der InsertDocu-Server .....	41
4	Performancemessungen .....	43
4.1	Messumgebung .....	43
4.2	Messaufbau .....	43
4.2.1	Dokumente .....	43
4.2.2	Queries .....	44
4.2.3	Server pro Komponente .....	45
4.2.4	Einfügeklient .....	46
4.2.5	Leseklient .....	46
4.3	Vergleich der Konflikttester .....	47
4.3.1	Konflikttest mit Verteilung der vollständigen Signaturen .....	47
4.3.2	Konflikttest mit Verteilung von Signaturteilen .....	48
4.3.3	Analyse .....	49
4.4	Messergebnisse mit einer Komponente .....	50
4.5	Messergebnisse mit zwei Komponenten .....	51
4.6	Messergebnisse mit vier Komponenten .....	53
4.7	Messergebnisse mit acht Komponenten .....	55
4.8	Vergleich der Antwortzeiten und der Durchsätze .....	56
4.8.1	Speedup von einer auf zwei Komponenten .....	56
4.8.2	Speedup von einer auf vier Komponenten .....	57
4.8.3	Speedup von einer auf acht Komponenten .....	58
4.8.4	Vergleich der maximalen Durchsätze .....	58
4.9	Messungen bei Hinzufügen von leeren Komponenten .....	59
4.10	Analyse der CPU-Zeiten der Server .....	61
5	Zusammenfassung .....	63
6	Ausblick .....	65
	Literaturverzeichnis .....	67
	Anhang A: Aufgabenstellung .....	69
	Anhang B: Messresultate im Detail .....	73
B.1	Resultate mit einer Komponente .....	73
B.2	Resultate mit zwei Komponenten .....	79
B.3	Resultate mit 4 Komponenten .....	87
B.4	Resultate mit 8 Komponenten .....	97
B.5	Resultate bei Hinzufügen von leeren Komponenten .....	110
	Anhang C: Startparameter der Server .....	115
C.1	Server DistrConfTest (Verteilung der vollständigen Signaturen) .....	115

---

C.2	Server DistrConfTest (2) (Verteilung der Signaturteile) .....	115
C.3	Server Forwarder (Verteilung der vollständigen Signaturen) .....	115
C.4	Server Forwarder (2) (Verteilung der Signaturteile) .....	115
C.5	Server GenerateSig .....	115
C.6	Server GetMaxDocId .....	116
C.7	Server GetTermId .....	116
C.8	Server GetTID .....	116
C.9	Server InsertDocu .....	116
C.10	Server MInsertDesc .....	116
C.11	Server MInsertDocu .....	116
C.12	Server MRetrieveDesc .....	117
C.13	Server MRetrieveDocu .....	117
C.14	Server RetrieveDocu .....	117
Anhang D: Startparameter der verwendeten Klienten .....		119
D.1	Klient changecc .....	119
D.2	Einfügeklient fclientsync .....	119
D.3	Leseklient fretclientsync .....	119
Anhang E: Beispiel eines Tuxedo-Konfigurationsfile .....		121
E.1	Ubb-File für zwei Komponenten und zwei Konflikttester .....	121
Anhang F: Datenbankschema .....		125
F.1	Tabellen .....	125
F.2	Primärschlüssel .....	126
Anhang G: Verteilung der Daten .....		127
G.1	2 Komponenten: .....	127
G.2	4 Komponenten: .....	127
G.3	8 Komponenten: .....	127
Anhang H: Englische Stopwörter .....		129