



FernUniversität in Hagen

– Fakultät für Mathematik und Informatik –
Lehrgebiet Datenbanken und Informationssysteme

Analyse und Optimierung der Webseite des Wedekind Projektes

Abschlussarbeit zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

vorgelegt von

Marco Galster

Matrikelnummer: 8335710

Referentin : Prof. Dr. Uta Störl

Betreuer : Sebastian Bruchhaus

ERKLÄRUNG

Ich erkläre, dass ich die Expose selbstständig und ohne unzulässige Inanspruchnahme Dritter verfasst habe.

Ich habe dabei nur die angegebenen Quellen und Hilfsmittel verwendet und die aus diesen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht.

Die Versicherung selbstständiger Arbeit gilt auch für enthaltene Zeichnungen, Skizzen oder graphische Darstellungen.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form weder derselben noch einer anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Mit der Abgabe der elektronischen Fassung der endgültigen Version der Arbeit nehme ich zur Kenntnis, dass diese mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate geprüft werden kann und ausschließlich für Prüfungszwecke gespeichert wird.

Hagen, 01. Oktober 2023

Marco Galster

INHALTSVERZEICHNIS

1	Expose	1
1.1	Ausgangslage	1
1.2	Ziel	2
1.3	Aktueller Forschungsstand	2
1.4	Vorgehen bei der Umsetzung	5
1.5	Vorläufige Gliederung der Abschlussarbeit	7
	Literatur	8

ABBILDUNGSVERZEICHNIS

Abbildung 1.1	Ablauf einer Web-Anfrage	5
---------------	------------------------------------	---

EXPOSE

1.1 AUSGANGSLAGE

Die Editions- und Forschungsstelle Frank Wedekind (EFFW) wurde 1987 in der Hochschule Darmstadt gegründet. Ihr Intention ist es, den lange vernachlässigten Autor der europäischen Moderne in die öffentliche Aufmerksamkeit zu bringen. Die Publikation der »Kritischen Studienausgabe der Werke Frank Wedekinds. Darmstädter Ausgabe« im Verlag Jürgen Häuser wurde 1994 direkt nach der Erschließung der Wedekind-Nachlässe in Aarau, Lenzburg und München begonnen und im Jahre 2013 abgeschlossen (8 Bände in 15 Teilbänden, jetzt in Wallstein Verlag). Die EFFW ist im Sommer 2015 an die Johannes Gutenberg-Universität Mainz umgezogen.

Da Frank Wedekind heute zu einen der bahnbrechenden Autoren der literarischen Moderne zählt, aber bisher sehr wenig erforscht wurde, soll sich dies nun Ändern. Die nationalen und internationalen Korrespondenzen von und an Wedekind zeigen eine starke Vernetzung in der europäischen Avantgarde. Dies zeigt das die Wissenschaft um die Korrespondenzen von Wedekind eine immer größere Rolle spielen. Aktuell sind lediglich 710 der 3200 bekannten Korrespondenzstücke veröffentlicht worden.

Um jenes zu verändern entstand das Projekt »Edition der Korrespondenz Frank Wedekind als Online-Volltextdatenbank« [Mar23], welches bei der EFFW angesiedelt ist und als Kooperationsprojekt an der Johannes Gutenberg-Universität Mainz, der Hochschule Darmstadt und der Fernuni Hagen umgesetzt und durch die Deutsch Forschungsgemeinschaft (Bonn) gefördert wird.

Hierbei wurden sämtliche bislang bekannten Korrespondenzen in die Online-Edition überführt. Diese beinhalteten substantiell das literarhistorische und kulturgeschichtliche Wissen über die Kultur zwischen 1880 und 1918, indem das überlieferte Material zum einen transkribiert editiert und zum anderen editionswissenschaftlich kommentiert wurde. Inhaltlich erschlossen zusätzliche Kommentare den historischen Kontext.

Hierfür entstand das Pilotprojekt der Online-Volltextdatenbank für Briefe von und an Frank Wedekind, welches 2015 als Beta-Version freigeschalten wurde. Dieses Projekt kann aktuell unter <http://briefedition.wedekind.h-da.de> eingesehen werden.

Die benutzerfreundliche Erfassung und Annotation der Briefe, ist eines der Hauptziele der konzeptionierten technischen Architektur. Aus diesem Grund, wurde die Präsentation-, Recherche- und Erstellungsebene vollständig webbasiert umgesetzt. Die Briefe selbst, sind im etablierten TEI-Format gespeichert. Dies muss von den Editoren und Editorinnen nicht selbst eingepflegt werden, sondern kann über einen entstanden WYSIWYG-Editor di-

rekt eingegeben werden, welcher dies bei der Speicherung in das TEI-Format umwandelt. Ebenfalls wurde hierbei auf eine modulare und unabhängige Architektur geachtet, wodurch die Komponenten im Nachgang auch von anderen Briefeditionen genutzt werden können.

1.2 ZIEL

Die aktuelle Umsetzung beinhaltet die bisher definierten Anforderungen vollständig, darunter fallen die Recherchemöglichkeiten, sowie auch die Eingabe und die Verarbeitung der Briefe. Ein größeres Problem hierbei ist die Performance der Oberfläche. Auf Grund der langen Abfragedauer des Datenbestandes leidet die Akzeptanz der Anwendung.

Das Ziel der Arbeit ist es, die Abfragedauer zu verringern, wodurch die Performance der Oberfläche signifikant verbessert wird.

Hierbei ist auch ein Vergleich mit anderen Technologien angedacht.

1.3 AKTUELLER FORSCHUNGSSTAND

Die Speicherverwaltung des PostgreSQL-Servers muss für Produktivsysteme angepasst werden [EH13, S. 34–38]. Hierunter fallen die *shared_buffers* die bei ca. 10 bis 25 Prozent des verfügbaren Arbeitsspeichers liegen sollten. Mit dieser Einstellung wird das häufige Schreiben des Buffers durch Änderungen von Daten und Indexen auf die Festplatte reduziert. Die Einstellung *temp_buffers* definiert wie groß der Speicher für temporäre Tabellen pro Verbindung maximal werden darf und sollte ebenfalls überprüft werden. Ein zu kleiner Wert bei großen temporären Tabellen führt zu einem signifikanten Leistungseinbruch, wenn die Tabellen nicht im Hauptspeicher, sondern in einer Datei ausgelagert werden.

Der *work_mem* definiert die Obergrenze des zur Verfügung gestellt Hauptspeichers pro Datenbankoperation wie effizientes Sortieren, Verknüpfen oder Filtern. Ebenso wird im Falle eines zu klein gewählten Speichers auf temporäre Dateien auf der Festplatte ausgewichen, was signifikanten Leistungseinbrüchen zur Folge haben kann. Die *maintenance_work_mem* wird bei Verwaltungsoperationen wie Änderungen und Erzeugungen von Datenbankobjekten als Obergrenze definiert. Die Wartungsaufgabe VACUUM, welche die fragmentierten Tabellen aufräumt und somit die Performance hebt, beachtet die Obergrenze ebenfalls.

Die Wartung des Datenbanksystems ist eine der wichtigsten Aufgaben und sollte regelmäßig durchgeführt werden, damit die Performance des Systems durch die Änderungen des Datenbestands nicht einbricht [EH13, S. 75]. Hierfür gibt es den VACUUM-Befehl, welcher entweder per Hand oder automatisch durch das Datenbanksystem ausgeführt werden soll. Für die automatische Ausführung kann der maximal verwendete Speicher über die Einstellung *autovacuum_work_mem* gesondert definiert werden [Posa]. Neben dem Aufräumen durch VACUUM, sollten auch die Planerstatistiken mit ANALYZE [EH13, S. 83] aktuell gehalten werden, damit die Anfragen durch

den Planer richtig optimiert werden können. Für beide Wartungsaufgaben gibt es den Autovacuum-Dienst, dieser sollte aktiv und richtig konfiguriert sein.

Mit dem Tool *pgFouine* [EH13, S. 155] können die Logs des PostgreSQL Server analysiert und auf Probleme hin untersucht werden. Hiermit können sehr einfach die häufigsten bzw. langsamsten Anfragen ermittelt werden.

Für weitere Optimierungen sollen werden anschließend die Anfragen einzeln überprüft. Hierfür ist es sinnvoll die Ausführungspläne der Abfrage zu analysieren [EH13, S. 252], die verschiedenen Plantypen und ihre Kosten zu kennen, sowie die angegebenen Werte für die Plankosten zu verstehen [DNB21, S. 24–30]. Besonderes Augenmerk gilt dem Vergleichen des tatsächlich ausgeführten mit dem ursprünglichen Plan [EH13, S. 254]. Eine der wichtigsten Kennzeichen hierbei ist, ob die Zeilenschätzung akkurat war, größere Abweichungen weisen häufig auf veraltete Statistiken hin.

Um die Abfragen selbst zu optimieren, gibt es ein Vorgehen über mehrere Schritte [DNB21, S. 304–308]. Zuerst wird Unterschieden, ob es sich um eine *Kurze* oder eine *Lange* Abfrage handelt. Im Falle einer *Kurzen* Abfrage, werden zuerst die Abfragekriterien überprüft. Sollte dies zu keiner Verbesserung führen, werden die Indexe geprüft. Ist dies ebenso erfolglos, wird die Abfrage nochmals genauer analysiert und so umgestellt, dass die restriktivste Einschränkung zuerst zutrifft. Bei einer *Langen* Abfrage soll überprüft werden, ob es sinnvoll ist, das Ergebnis in einer Tabelle zu speichern und bei Änderungen zu aktualisieren. Wenn dies nicht möglich ist, sollten die folgenden Schritte durchgeführt werden. Zuerst wird der restriktivste Join gesucht und überprüft, ob dieser als Erstes ausgeführt wird. Anschließend fügt man weitere Joins hinzu und prüft die Ausführungszeit und die Abfragepläne. Als Nächstes wird sich vergewissert, ob große Tabellen nicht mehrfach durchsucht worden sind. Bei Gruppierungen ist noch zu prüfen, ob diese früher durchgeführt werden können, um die Abfragemenge zu verringern.

Bei *Langen* Abfragen ist die Abhandlung »Optimizing Iceberg Queries with Complex Joins« [WRY17] ein zusätzlicher Ratgeber, um die Performance zu steigern.

Des Weiteren können über das Modul `pg_stat_statements` Statistiken der Aufrufe die an den Server gestellt wurden, ermittelt werden [Posb]. Hierbei können die am häufigsten Aufgerufenen und die Anfragen mit der längsten Ausführungszeit ermittelt werden.

Die *Java Persistence API (JPA)* wird als First-Level-Cache in Java-EE-Anwendung verwendet, hier nehmen die Objekte einen von vier Zuständen ein [MW12, S. 57]. Im Zustand *Transient* sind die Objekt erzeugt, aber noch nicht in den Cache überführt worden. Wenn diese in den Cache überführt worden sind, nehmen sie den Zustand *Verwaltet* ein. Ist das Objekt aus dem Cache und der Datenbank entfernt worden, nimmt es den Zustand *Gelöscht* an. *Losgelöst* ist der letzte Zustand, bei dem das Objekt aus dem Cache entfernt worden ist, aber nicht aus der Datenbank.

Eine Menge von Objekten wird als *Persistenzkontext* bezeichnet. Solange die Objekte dem *Persistenzkontext* zugeordnet sind, also den Zustand *Veraltet* besitzen, werden diese auf Änderungen überwacht, um sie am Abschluss mit der Datenbank zu synchronisieren. In der Literatur wird hierzu der Begriff *Automatic Dirty Checking* verwendet [MW12, S. 61].

In den Java-EE-Anwendungen wird der *Persistenzkontext* für die Anfragen vom *Application-Server* bereitgestellt. Hierfür werden *Application-Server* wie *GlassFish* genutzt, um die Verwendung eines Pools von Datenbankverbindungen zu definieren [MW12, S. 68]. Dadurch kann die Anzahl der Verbindung geringer gehalten werden als die Anzahl der Benutzer die an der Anwendung arbeiten. Zusätzlich werden die Transaktionen über *Stateful Session-Bean (SFSB)* gehandhabt, welche automatisch vor dem Aufruf erzeugt und danach wieder gelöscht werden. Dies birgt allerdings den Nachteil, dass der *Persistenzkontext* sehr groß werden kann, wenn viele Entities in den *Persistenzkontext* geladen werden. Da dies häufig zu Speicher- und damit Performanz-Problemen [MW12, S. 79] führen kann, muss hier darauf geachtet werden, nicht mehr benötigte Entities aus dem *Persistenzkontext* zu lösen.

Zusätzlich kann im *JPA* ebenfalls noch der *Second Level Cache (L2-Cache)* aktiviert werden. Dieser steht jedem *Persistenzkontext* zur Verfügung und kann dadurch die Anzahl der Datenbankzugriffe deutlich reduzieren, was bei langsamen Datenbank-Anbindungen zu hohen Performance-Gewinnen führen kann [MW12, S. 171]. Gegen die Verwendung spricht, dass die Daten im *Second Level Cache* explizit über Änderungen informiert werden müssen, welche sonst beim nächsten Aufruf veraltete Werte liefern. Ebenfalls benötigt so ein Cache einen höheren Bedarf an Arbeitsspeicher, in dem die Daten parallel zur Datenbank bereitgestellt werden, daher ist die Benutzung nur problemlos bei Entities möglich, auf die meist lesend zugegriffen wird.

In der *OpenJPA*-Erweiterung für den L2-Cache, wird in *Objekt-Cache* (in *OpenJPA* als *DataCache* bezeichnet) und *Query-Cache* unterschieden. Über die Funktionen *find()* und *refresh()* oder einer *Query* werden die geladenen Entities in den Cache gebracht. Davon ausgenommen sind *Large Result Sets* (Abfragen die nicht alle Daten auf einmal laden), *Extent-Technologien* und *Queries*, die einzelne Attribute von Entities zurückliefern, aber nicht das Entity selbst. Hierbei kann genau gesteuert werden, welche Entity in den Cache abgelegt wird und welche nicht. Ebenfalls kann auf Klassenbasis der zugehörige Cache definiert werden, um eine bessere Last-Verteilung beim Zugriff zu ermöglichen [MW12, S. 314].

Im *Query-Cache* werden die Abfragen bzw. die Eigenschaften einer Abfrage und die zurückgelieferten Ids der Entities gespeichert. Bei einem erneuten Aufruf dieser Abfrage werden die referenzierten Objekte aus dem *Objekt-Cache* zurückgegeben. Bei veränderten referenzierten Entities wird der *Query-Cache* nicht genutzt und die betroffenen Abfragen werden unverzüglich aus dem *Query-Cache* entfernt [MW12, S. 316].

Um zu prüfen, ob die Einstellungen sinnvoll gesetzt sind, kann in *OpenJPA* eine *Cache-Statistik* abgefragt werden. Mit dieser kann die Anzahl der Lese- und Schreibzugriffe im Cache überprüft werden, entsprechend die-

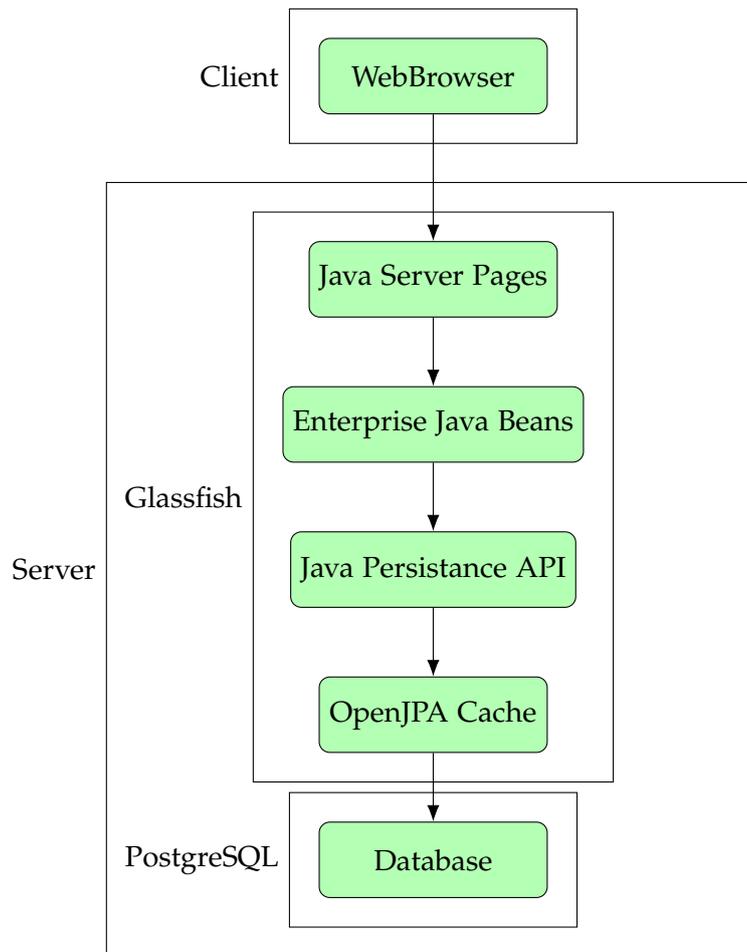


Abbildung 1.1: Ablauf einer Web-Anfrage

ser Auswertung sollten die Einstellungen an den Entities angepasst werden [Ibm].

1.4 VORGEHEN BEI DER UMSETZUNG

Durch eine Umfrage der Bediener und Entwickler, einer Performance-Messung in der Webseite und den Statistiken im PostgreSQL, sollen die größten Performance-Probleme in der Webseite ermittelt und der dazugehörigen Quellcode identifiziert werden. Für die Analyse und Optimierung der Abfragen sollen verschiedene Blickwinkel betrachtet werden.

Bei den einzelnen Abfragen muss zuerst ermittelt werden, in welchem Teil des Aufrufs die meiste Zeit aufgewendet wird, hierbei wird die Übertragung über das Netzwerk außer acht gelassen, da diese vom Standort und nicht direkt von der Anwendung abhängt. Ein geplantes Vorgehen ist hierbei die Überprüfung von »unten nach oben«, wie in 1.1 dargestellt.

Zuerst soll der Aufruf gegen die Datenbank geprüft und die Ausführungszeit sowie die Abfragepläne ermittelt und analysiert. Wenn sich hierbei größere Defizite erkennen lassen, werden die Abfragen direkt optimiert, indem

der Aufbau der Abfrage, die Abfragekriterien und die Verwendung der Indexe betrachtet und in Frage gestellt werden.

Anschließend erfolgt die Prüfung des OpenJPA-Caches mit den zugehörigen Statistiken. Bei diesen wird einerseits ermittelt, wie sinnvoll der aktuelle Einsatz des Caches für die unterschiedlichen Entitäten ist und andererseits ob es sinnvoll ist die aktuelle Nutzung zu minimieren oder ihn komplett zu entfernen. Ein Ersatz mit besserer Performance soll in diesem Fall ebenso untersucht werden.

Anschließend wird der Aufruf über die JPA betrachtet. Dies ist besonders wichtig, wenn die Abfragen dynamisch erzeugt werden und die SQL-Abfrage selbst nicht optimiert werden kann. In diesem Fall sollte nicht nur die reine Abfrage, sondern auch die Verwendung des Caches mit in Betracht gezogen werden.

Nun wird die EJB-Schicht überprüft und die aufgerufen Funktionen betrachtet, ob hier einzelne Funktionen zu viele Aufgaben übernehmen und dadurch schlecht optimiert werden können. Solche Funktionen sollten dupliziert werden und auf die jeweilige Aufgabe spezifisch zugeschnitten und optimiert werden.

Abschließend ist die JSP-Schicht zu betrachten, welche noch logische Anpassungen für die Optimierungen zulässt, wie das Einfügen von Paging. Damit kann die ermittelnde Datenmenge verringert werden, dies führt zu schnelleren Abfragen, weniger Daten die im Cache vorgehalten werden müssen, den Aufbau der Seite zu beschleunigen und damit auch die Datenmenge die an den Browser übermittelt wird zu reduzieren.

Zeitgleich werden der PostgreSQL sowie der Server selbst untersucht und die Einstellungen überprüft. Hierzu gehören die Größen der Speicher und die Wartungsaufgaben des Datenbanksystems. In diesem Zuge werden auch die Log-Dateien vom PostgreSQL, unter Zuhilfenahme von pgFouine untersucht und auf Probleme und Unregelmässigkeiten überprüft.

1.5 VORLÄUFIGE GLIEDERUNG DER ABSCHLUSSARBEIT

1. Einleitung
2. Grundlagen
3. Konzept
 - 3.1 Aufbau der Umfrage
 - 3.2 Allgemeine Betrachtung des Systems
 - 3.3 Das Vorgehen der Optimierungen
 - 3.4 Aktueller Aufbau der Software
 - 3.5 Vergleich mit anderen Technologien
4. Performance-Untersuchung
 - 4.1 Auswertung der Umfrage
 - 4.2 Einbau und Aktivieren von Performance-Messung
 - 4.3 Statistiken im PostgreSQL auswerten
 - 4.4 Überprüfung des PostgreSQL und Servers
5. Optimierung
 - 5.1 Ermittlung der Performance-Probleme
 - 5.2 Analyse der Abfragen
 - 5.3 Optimierungen der Abfragen
 - 5.4 Anpassung der Konfiguration
6. Evaluierung
 - 6.1 Befragung der Benutzer und Entwickler
 - 6.2 Erneute Laufzeitanalyse starten
 - 6.3 Statistiken im PostgreSQL auswerten
 - 6.4 Vergleich der Ergebnisse vor und nach der Optimierung
7. Zusammenfassung und Ausblick

LITERATUR

- [Posa] 2023. URL: <https://postgrespro.com/docs/postgresql/14/runtime-config-resource> (besucht am 27. 12. 2023).
- [Posb] 2023. URL: <https://www.postgresql.org/docs/8.4/pgstatstatements.html> (besucht am 27. 12. 2023).
- [Ibm] 2023. URL: <https://www.ibm.com/docs/de/was/8.5.5?topic=applications-configuring-openjpa-caching-improve-performance> (besucht am 24. 09. 2023).
- [DNB21] Henrietta Dombrovskaya, Boris Novikov und Anna Bailliekova. *PostgreSQL Query Optimization - The Ultimate Guide to Building Efficient Queries*. Berkeley, CA: Apress, 2021. ISBN: 978-1-4842-6885-8. DOI: 10.1007/978-1-4842-6885-8. eprint: <https://link.springer.com/book/10.1007/978-1-4842-6885-8>. URL: <https://link.springer.com/book/10.1007/978-1-4842-6885-8>.
- [EH13] Peter Eisentraut und Bernd Helmle. *PostgreSQL-Administration* -. Köln: O'Reilly Germany, 2013. ISBN: 978-3-868-99362-2.
- [Mar23] AK Prof. Dr. Ariane Martin. *Frank Wedekind | FB 05 - AK Prof. Dr. Ariane Martin*. 2023. URL: <https://www.martin.germanistik.uni-mainz.de/forschung/frank-wedekind/> (besucht am 24. 09. 2023).
- [MW12] Bernd Müller und Harald Wehr. *Java Persistence API 2*. München: Carl Hanser Verlag GmbH & Co. KG, 2012. DOI: 10.3139/9783446431294. eprint: <https://www.hanser-elibrary.com/doi/pdf/10.3139/9783446431294>. URL: <https://www.hanser-elibrary.com/doi/abs/10.3139/9783446431294>.
- [WRY17] Brett Walenz, Sudeepa Roy und Jun Yang. "Optimizing Iceberg Queries with Complex Joins". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. Chicago, Illinois, USA: Association for Computing Machinery, 2017, 1243–1258. ISBN: 9781450341974. DOI: 10.1145/3035918.3064053. URL: <https://doi.org/10.1145/3035918.3064053>.