

Projekt NoSQL

Author: Lukas Lieb

Daniel Schmid

Matthias Eiholzer

E-Mail: lukas.lieb@stud.hslu.ch

daniel.schmid@stud.hslu.ch

matthias.eiholzer@stud.hslu.ch

Erstellungsdatum: 21. Dezember 2016

Inhaltsverzeichnis

1	Einführung	5
1.1	Aufgabenstellung	5
1.2	Datenbankauswahl	6
2	Datenmanagement	7
3	Datenmodellierung	9
4	Datenbanksprachen	12
5	Konsistenzsicherung	15
6	Systemarchitektur	17
7	Schlussfolgerungen	19

Abbildungsverzeichnis

1	Visualisierung des Anwendungsfalls	7
2	Gui für den Benutzer	8
3	ER-Diagramm zur Uni-DB Kaufmann (2016)	9
4	ER-Diagramm zur Uni-DB der NoSQL Datenbank	10
5	Query Resultat in MySQL	13
6	Query Resultat MongoDB	14

Tabellenverzeichnis

1 Einführung

1.1 Aufgabenstellung

Ziel dieser Arbeit ist es, die Uni-DB, welche auf einer SQL Server liegt, in eine NoSQL-Datenbank zu implementieren. Dazu Wird das ER-Schema der Uni-DB so umgezeichnet werden, dass es in der gewählten NoSQL-Datenbank abgebildet werden kann. Dieses Schema wird dann in der entsprechend gewählten Datenbank umgesetzt. Dazu migriert man die Daten aus der SQL- in die NoSQL-Datenbank. Zusätzlich wird dem Benutzer eine Möglichkeit geboten werden, mithilfe eines GUI folgendes SQL-Query auf der MongoDB abzufragen:

```
select ProfessorName , AnzahlStudenten , SummeSWS
from (
select p.Name as ProfessorName , count(s.MatrNr)
as AnzahlStudenten
from Professoren p
join Vorlesungen v on v.gelesenVon = p.PersNr
join hoeren h on h.VorlNr = v.VorlNr
join Studenten s on s.MatrNr = h.MatrNr
group by p.Name
) A
join
(
select p.Name as ProfessorName , sum(SWS)
as summeSWS
from Professoren p
join Vorlesungen v on v.gelesenVon = p.PersNr
group by p.Name
) B using(ProfessorName)
```

1.2 Datenbankauswahl

Bei der Auswahl der Datenbank haben wir uns für Die MongoDB entschieden. Entscheidend waren dabei folgende Gründe:

- Geeignet um unser Problem zu lösen.
- Wird in der Praxis eingesetzt
- Bekannt
- Gute Dokumentation
- Kostenlos
- Unterstützung durch Forenuser
- Erste Erfahrung vorhanden

Die MongoDB ist eine Datenbank, welche sich grundlegend im CAP Theorem in der Spalte CP aufhält. Dies bedeutet, dass die Konsistenz (Consistency) gewährleistet ist. Dies wird realisiert, indem in einem System ein Knoten als primäres Mitglied fungiert, und alle Anfragen über diesen Knoten abgearbeitet werden. Zusätzlich wird die Partition-Tolerance gewährleistet. Dies funktioniert laut Vertreiber, indem wenn der Primäre Knoten ausfällt, automatisch ein sekundärer Knoten als neuer Primärknoten definiert wird und somit das System wieder funktioniert. Ebenfalls soll es möglich sein, auf Kosten der Konsistenz (Consistency) die Verfügbarkeit (Availability) zu erhöhen, indem man in den Einstellungen die Konfiguration vornimmt, dass Daten nicht bloss über den Primär-Knoten, sondern auch über Sekundärknoten bezogen werden können.

2 Datenmanagement

Zwei Akteure wirken auf die Datenbank ein. Einerseits der Benutzer, welcher die Informationen abfragen kann. Andererseits der DB-Administrator, welcher die Daten abfragen, aber auch verändern und neue hinzufügen kann. Diese Anwendungsfälle sind in der Abbildung 1 abgebildet.

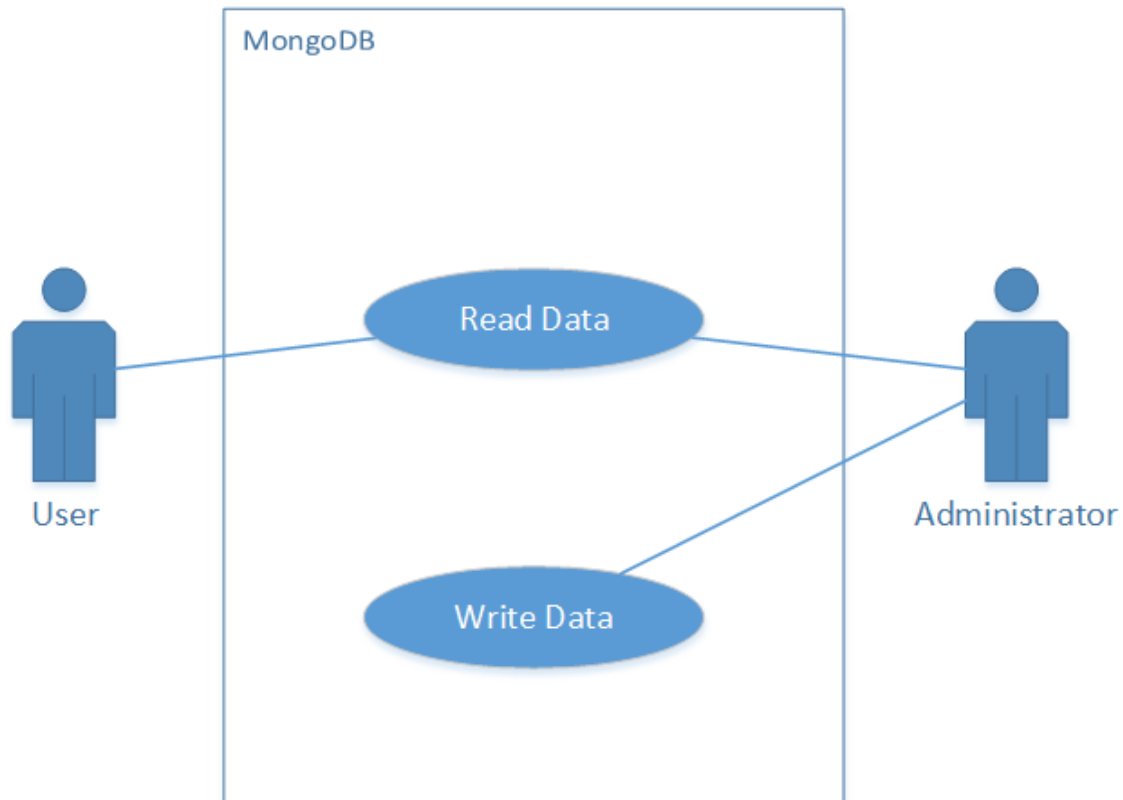
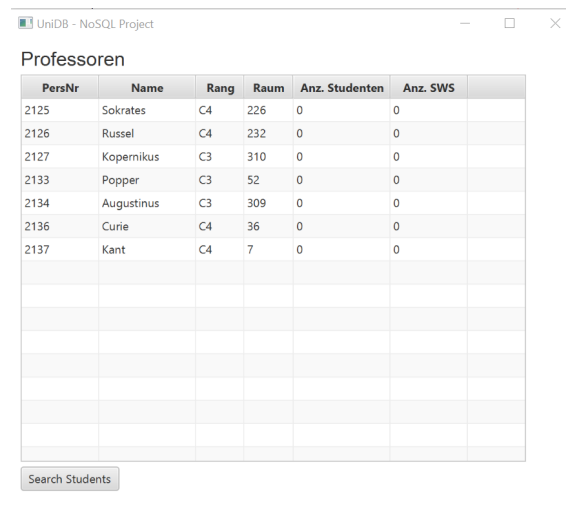


Abbildung 1: Visualisierung des Anwendungsfalls

Die Daten, welche für die MongoDB verwendet werden, wurden aus der Uni-DB, welche für einige Testaufgaben im Modul DMG bereits verwendet wurde, migriert. Um diese Migration durchzuführen wurde ein Java Programm geschrieben, welches in Kapitel 4 näher erläutert wird. Der Strukturierte Aufbau der Datenbank wird in Kapitel 3 beschrieben.

Der Benutzer hat die Möglichkeit, über ein GUI Daten auf der Datenbank abzufragen:



UniDB - NoSQL Project

Professoren

PersNr	Name	Rang	Raum	Anz. Studenten	Anz. SWS
2125	Sokrates	C4	226	0	0
2126	Russel	C4	232	0	0
2127	Kopernikus	C3	310	0	0
2133	Popper	C3	52	0	0
2134	Augustinus	C3	309	0	0
2136	Curie	C4	36	0	0
2137	Kant	C4	7	0	0

Search Students

Abbildung 2: Gui für den Benutzer

Über dieses GUI sind alle Professoren ersichtlich, welche in der Datenbank eingetragen sind. Mithilfe des SearchStudents Knopf kann der Benutzer abfragen, wie viele Studenten der entsprechende Professor in seinen Vorlesungen hat, und wie viele SWS Punkte er unterrichtet.

3 Datenmodellierung

MongoDB ist eine dokumentorientierte Datenbank. Dabei werden die Daten in JSON ähnlichen Dokumenten verwaltet. Das bedeutet, dass die Daten nicht relational verwaltet werden. So kann zum Beispiel ein Tupel einer relationalen Datenbank als Dokument in der dokumentorientierten Datenbank abgebildet werden. Die Attribute und die dazugehörigen Werte werden dabei in Schlüssel-Wert Paare abgebildet.

Unserem Projekt liegt das ER-Schema aus Abbildung 3 zugrunde. Da MongoDB eine NoSQL und keine relationale Datenbank ist, kann das ER-Schema nicht direkt in dieser Form in der Datenbank abgebildet werden.

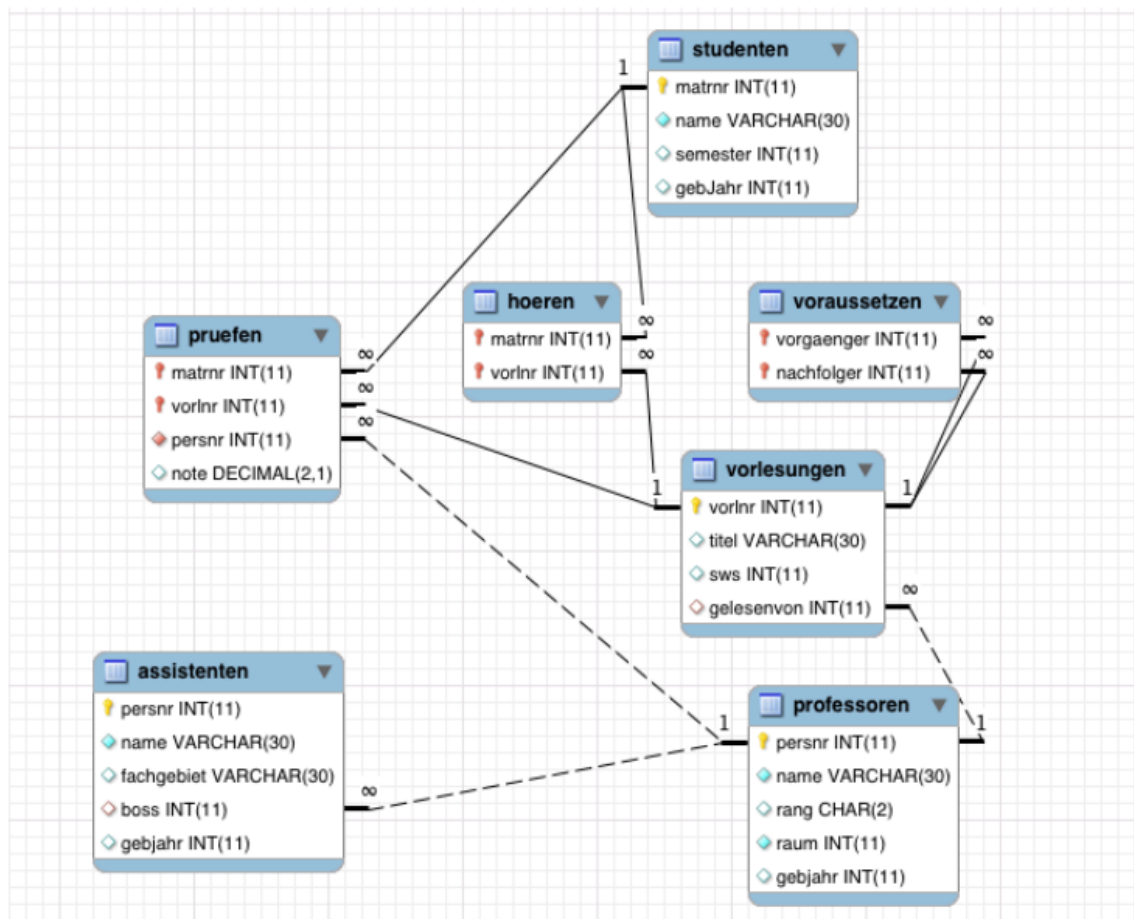


Abbildung 3: ER-Diagramm zur Uni-DB Kaufmann (2016)

Das Schema unserer NoSQL Datenbank sieht wie folgt aus:

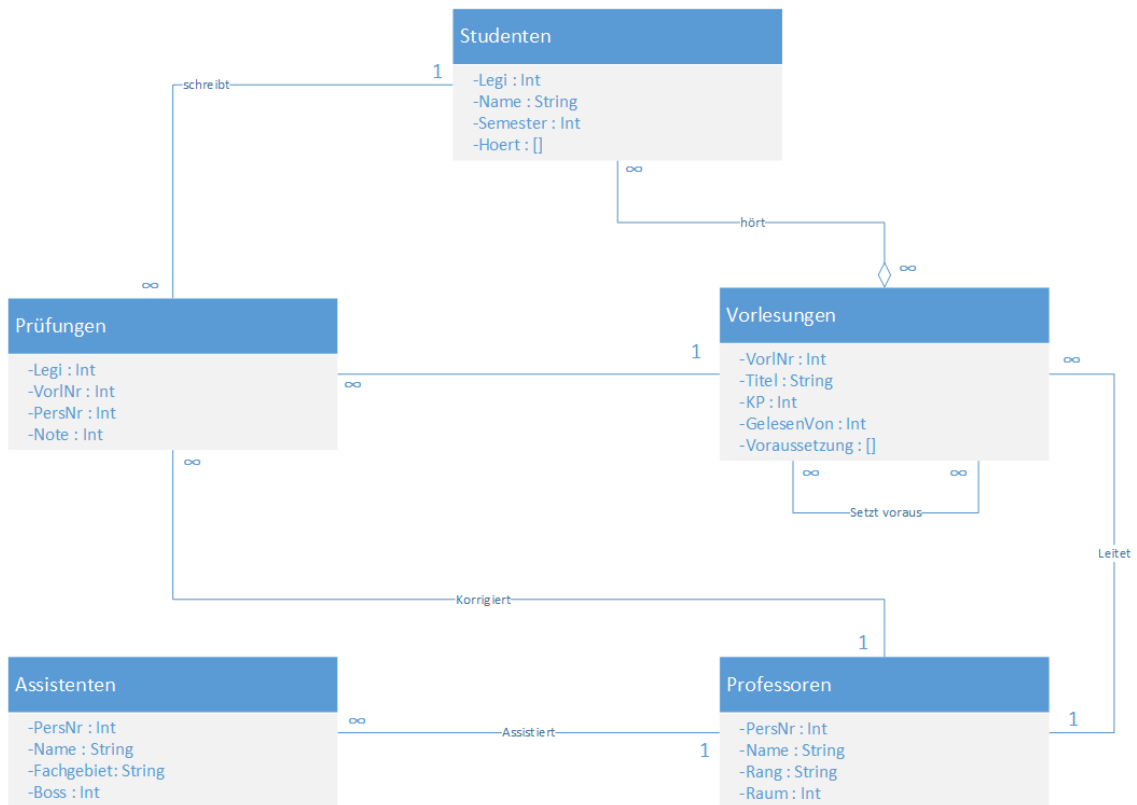


Abbildung 4: ER-Diagramm zur Uni-DB der NoSQL Datenbank

Eine Komposition bedeutet, dass z.B in der NoSQL Datenbank als “teil von” abgebildet wird. Eine Aggregation bedeutet, dass diese in Beziehung als Referenz abgebildet wird.

Im ER-Diagramm wird zum Beispiel die Komplex-Komplexe-Beziehung zwischen Vorlesungen und Studenten als eigene Tabelle abgebildet. In unserer Datenbank wird diese Beziehung in zwei einfach-komplexe Beziehung abgebildet. Somit besitzen Studenten eine Einfach-Komplexe Beziehung zu den Vorlesungen. Diese Beziehung wird im JSON -Format im Feld Hören ersichtlich. Die vom Studenten besuchten Vorlesungen werden als Vorlesungsnummer in einem Array referenziert.

Anbei noch ein Beispiel wie ein Student in der aufgesetzten MongoDB nach vorgegebenen ER-Diagramm abgespeichert ist:

```
{  
  "Legi": 25403,  
  "Name" : "Jonas",  
  "Semester" : 12,  
  "Hören" : [5032, 1910],  
}
```

4 Datenbanksprachen

MongoDB bietet verschiedene Möglichkeiten um Daten hinzuzufügen oder zu manipulieren. So können Dokumente beispielsweise über die Kommandozeile hinzugefügt werden. Zudem gibt es für viele Programmiersprachen einen Treiber, der in das jeweilige Projekt eingebunden werden kann.

Da für diese Arbeit ein Programm in Java geschrieben wurde, kam der aktuelle mongo-java-driver in der Version 3.4.0 zum Einsatz. Allerdings sind die Kommandos für die Interaktion mit den Daten nicht sehr unterschiedlich zum normalen Kommandofenster. Nach der Wahl der Datenbank kann im Java Treiber die entsprechende Collection ausgewählt und dieser anschliessend Dokumente hinzugefügt werden:

```
MongoClient mongo = new MongoClient( "localhost" , 27017 );
MongoDatabase database = mongo.getDatabase("unldb");
MongoCollection<Document> stud = database.getCollection("studenten");
stud.drop();
studs.insertOne(new Document("Legi", 25403)
    .append("Name", "Jonas")
    .append("Semester", 12)
    .append("Hoeren", 5022));
mongo.close();
```

So wurde für die Testdaten ein Javaprogramm erstellt, dass alle Daten der UniDB in entsprechende Collections der MongoDB schreibt.

Für die Erwähnte Query, die in der Einführung bereits erwähnt wurde, wurde eine separate Funktion geschrieben, die die Daten aus der Datenbank holt und aneinanderhängt. Der Join wird bei den zwei for-Schleifen ausgeführt. Darin wird für jeden Professor nachgeschlagen, welche Vorlesungen dieser hält. Anhand dieser wird die Anzahl SWS Punkte ermittelt, die der Professor unterrichtet. In der zweiten for-Schleife wird zusätzlich gezählt, wie viele Studenten dieser Vorlesungen zuhören. Die Selektion wird hier bei der Setzung der Anzahl SWS und Studenten zur List aller Professoren ausgeführt (p.setAnzSWS(swsCounter) und p.setAnzStud(studCounter)).

```
MongoClient mongo = new MongoClient( "localhost" , 27017 );
MongoDatabase database = mongo.getDatabase("unldb");
MongoCollection<Document> prof = database.getCollection("professoren");
MongoCollection<Document> vorl = database.getCollection("vorlesungen");
MongoCollection<Document> stud = database.getCollection("studenten");
```

```

for (Professor p: professoren){
    int swsCounter = 0;
    int studCounter = 0;

    for(Document dVorl: vorl.find(eq("GelesenVon", p.getPersNr()))){
        swsCounter += Integer.parseInt(dVorl.get("SWS").toString());
        for(Document dStud: stud.find(eq("Hoeren", dVorl.get("VorlNr")))){
            studCounter ++;
        }
    }
    p.setAnzSWS(swsCounter);
    p.setAnzStud(studCounter);
}

```

Wenn die Query in MySQL ausgeführt wird, werden als Resultat die Anzahl Studenten und die Summe der SWS für jeden Professor angezeigt:

ProfessorName	AnzahlStudenten	SummeSWS
Augustinus	2	2
Kant	4	8
Popper	1	2
Russel	2	8
Sokrates	4	10

Abbildung 5: Query Resultat in MySQL

Beim Javacode wird diese eine Tabelle angezeigt und mit einem Click auf den Search Students Knopf wird besagte Funktion oben ausgeführt und die Tabelle aktualisiert.

The image displays two screenshots of a web application titled 'UniDB - NoSQL Project'. Both screenshots show a table titled 'Professoren' with the following columns: PersNr, Name, Rang, Raum, Anz. Studenten, and Anz. SWS. Below the table is a 'Search Students' button.

Left Screenshot (Initial State):

PersNr	Name	Rang	Raum	Anz. Studenten	Anz. SWS
2125	Sokrates	C4	226	0	0
2126	Russel	C4	232	0	0
2127	Kopernikus	C3	310	0	0
2133	Popper	C3	52	0	0
2134	Augustinus	C3	309	0	0
2136	Curie	C4	36	0	0
2137	Kant	C4	7	0	0

Right Screenshot (After Search):

PersNr	Name	Rang	Raum	Anz. Studenten	Anz. SWS
2125	Sokrates	C4	226	4	10
2126	Russel	C4	232	2	8
2127	Kopernikus	C3	310	0	0
2133	Popper	C3	52	1	2
2134	Augustinus	C3	309	2	2
2136	Curie	C4	36	0	0
2137	Kant	C4	7	4	8

Abbildung 6: Query Resultat MongoDB

5 Konsistenzsicherung

Bei relationalen Datenbanken bietet die Datenbank Mechanismen zur Konsistenzsicherung an. Dies wird durch verschiedene Eigenschaften im RDBMS erreicht. Jeder Tabelle liegt ein Schema zu Grunde. In diesem wird definiert, wie die Daten strukturiert sein müssen. Zum Beispiel wird festgelegt, welches der Primärschlüssel ist, ob Felder “null” sein dürfen, oder ob Tupel gelöscht werden dürfen, wenn noch Referenzen darauf zeigen. MongoDB ist Schema frei. Das bedeutet, dass sich je zwei Dokumente in einer Sammlung komplett in ihrer Struktur voneinander unterscheiden können. Des Weiteren garantiert MongoDB keine Integrität der Referenzen. Dieser Umstand wird klar, wenn man vergleicht, welchen Regeln die relationale Datenbank und die MongoDB unterliegen. RDBMS unterliegen den ACID Regeln.

- Atomar: Die gesamte Transaktion wird ausgeführt oder die Transaktion wird rückgängig gemacht.
- Consistent: Nach jeder Transaktion muss die Datenbank widerspruchsfrei sein.
- Isolatet: Bei einer Transaktion dürfen keine Seiteneffekte auftreten. Dies garantiert, dass ein Mehrbenutzerbetrieb möglich ist.
- Dauerhaft: Die Daten werden sicher gespeichert, auch bei Systemabstürzen. Bei einem Systemabsturz wird ein recovery durchgeführt, so dass danach die Datenbank wieder in einem konsistenten Zustand ist.

Die in unserem Projekt eingesetzte MongoDB unterliegt den BASE Regeln.

- Basically Available: Die Datenbank sollte meistens laufen.
- Eventually Consistent: Die Konsistenz der Daten wird nicht unmittelbar nach der Operation gewährleistet. Sie kann verzögert eintreten.

Da MongoDB keine Transaktionen (eine Folge von Operationen, die Atomar ausgeführt werden)) unterstützt, muss diese Funktionalität auf der Anwendungsebene implementiert werden. Dies ist beim Einsatz eines RDBMS nicht notwendig, da dieses Transaktionen unterstützt. In unserem Fall werden keine Transaktionen

verwendet, da nur lesend auf die Daten zugegriffen werden. Da nur gelesen wird, und keine Daten geändert oder hinzugefügt/entfernt werden, kann die Datenbank durch Operationen nicht in einen inkonsistenten Zustand überführt werden. Deswegen kann auch parallel auf die Daten zugegriffen werden, ohne die Konsistenz zu verlieren. Sollen zu einem späteren Zeitpunkt zur Laufzeit der Anwendung Daten geändert oder hinzugefügt/entfernt werden, so muss eine Konsistenzsicherung auf der Anwendungsebene implementiert werden.

6 Systemarchitektur

MongoDB besteht aus drei verschiedenen Servern. Den Routern, Config Servern und den Replica Sets. Jeder dieser drei Servertypen hat eine bestimmte Aufgabe. Der Client sendet seine Operation an den Router. Der Router wiederum leitet die Operation an die zuständigen Replica Sets weiter. Die Replica Sets wiederum speichern die zu verarbeitenden Dokumente. Damit der Router weiss, an welche Replica Sets er die Anfrage weiterleiten muss, stellt er wiederum eine Anfrage an die Config Server. Diese Antworten entsprechend. Die Config Server enthalten also die Metadaten über die Dokumentenverteilung in den Replica Sets. In der Abbildung ?? ist die Architektur von MongoDB visualisiert. In der in diesem Projekt eingesetzte MongoDB Instanz, laufen alle drei Server auf demselben physischen Rechner. Dies wurde so gewählt, da die Sammlungen an Dokumenten klein ist. Des Weiteren wird auch nur sporadisch auf die Daten zugegriffen. Sollte sich in Zukunft einer dieser Parameter ändern, kann darüber nachgedacht werden, die Server auf verschiedene physischen Server zu verteilen. Dies hätte auch den weiteren Vorteil, dass Parallel auf die Dokumente zugegriffen werden kann. Ist die Replica Set 1 mit der Verarbeitung einer Anfrage beschäftigt, kann währenddem die Replica Set 2 die nächste Anfrage bearbeiten, sofern es sich bei der Anfrage um die in ihr gespeicherten Dokumente handelt.

In RDBMS liegen die Tabellen in normalisierter Form vor. Daten die nicht funktional vom Primärschlüssel abhängen, werden als Fremdschlüssel referenziert. Die Referenzierung macht es schwierig, die Tabelle, in welcher das referenzierte Tupel abgelegt ist, auf einen anderen Server auszulagern. Denn bei jedem Zugriff auf die ausgelagerte Tabelle, wird die Abfrage der Daten verlangsamt, da zwischen den beiden Servern Kommunikation stattfindet. Dieses Problem hat MongoDB nicht, da die Daten nicht normalisiert vorliegen müssen. Bei MongoDB dürfen die bei RDBMS referenzierten Tupel als Aggregationen in die Dokumente umgesetzt werden. Dies vermindert die Anzahl der Referenzierungen und ermöglicht es damit, dass eine horizontale Skalierung einfacher zu bewerkstelligen ist, als bei einer relationalen Datenbank. Dies führt auch Nachteile mit sich. Jedes mal wenn ein referenziertes Tupel

als Aggregation in ein Dokument gespeichert wird, erhöht sich die Datenmenge. Des weiteren bedeutet es mehr Aufwand, ein solches aggregiertes Tupel zu ändern, da die Änderung bei allen Dokumenten gemacht werden muss, bei denen das referenzierte Tupel als Unterteil gespeichert wird. Dies im Gegensatz zur relationalen Datenbank. Bei dieser muss die Änderung nur an einem Ort stattfinden. MongoDB kennt keine Funktionlität zum Auflösen von Referenzen. Diese muss jeweils in der Applikation implementiert werden.

Da wie bereits erwähnt, alle Server auf einem physischen Rechner laufen, ist die Datenbank bei einem Systemausfall nicht mehr erreichbar. Bei einem Festplatten-ausfall kann es sogar passieren, dass die komplette Datenbank verloren geht, da es in der Replica Set nur einen Server gibt, und somit ein weiterer Server fehlt, der ein Replikat der Datenbank enthält.

7 Schlussfolgerungen

Literatur

Kaufmann, Michael., “Übung S1: SQL Grundlagen,” 2016.