Other Databases: Graph Database Movie Research

Benno Grimm, Anna-Lena Richert, Marcel Mertens & Anton Ochel

Was ist eine Graph Datenbank?

Graph Datenbank

> "A graph database is a database designed to treat the relationships between data as equally important to the data itself."
>
> - Neo4j

- Daten werden in Graphen gespeichert und dargestellt
- Es gibt Knoten und Linien: Knoten stellen Entitäten dar und Linien stellen Beziehungen dar
- Ein Knoten kann beliebig viele Eigenschaften haben
- Eine Beziehung hat immer eine Richtung, eine Art, einen Startknoten und einen Endknoten
    - Beziehungen können außerdem noch Eigenschaften haben
    - Jeder Knoten kann beliebig viele Beziehungen mit anderen Knoten haben

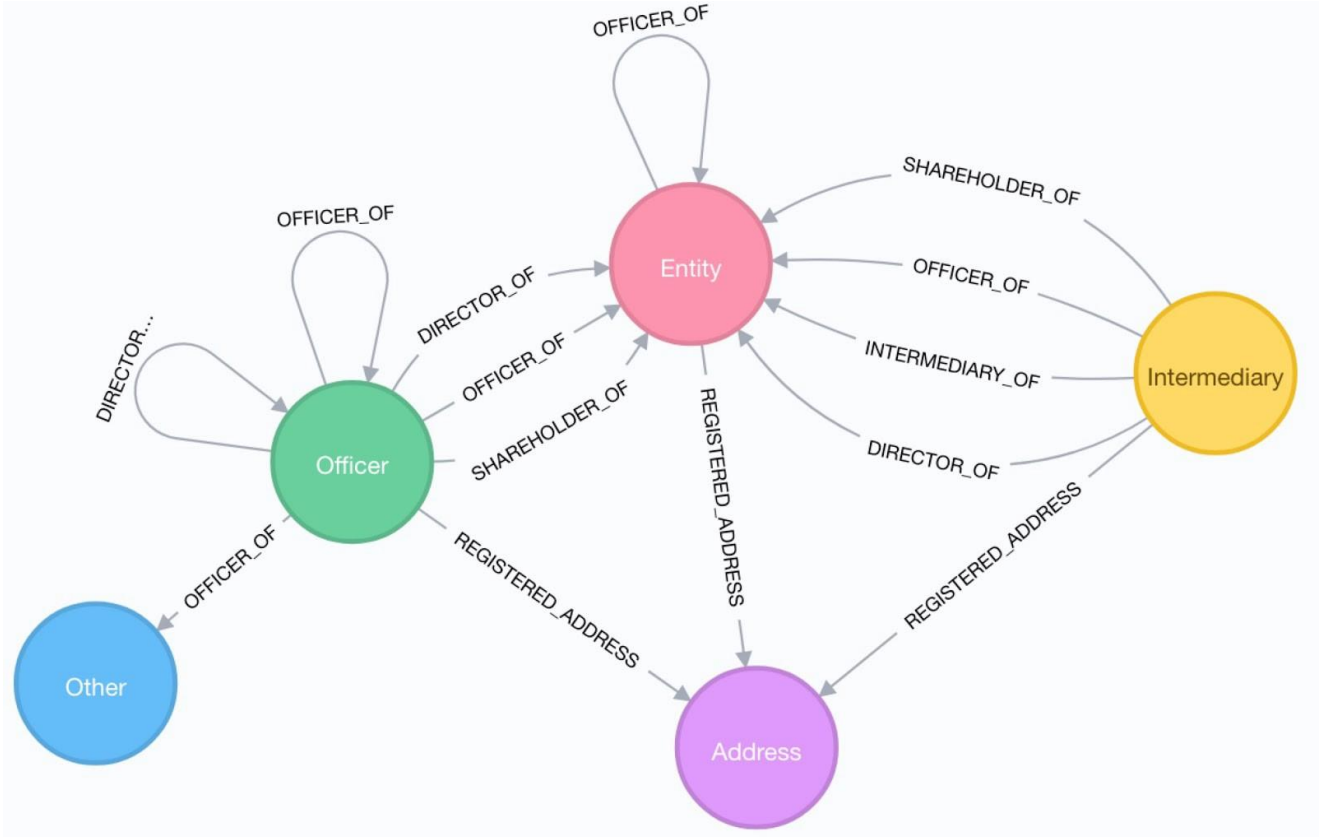

Abb.: Daten repräsentiert in einem Graph

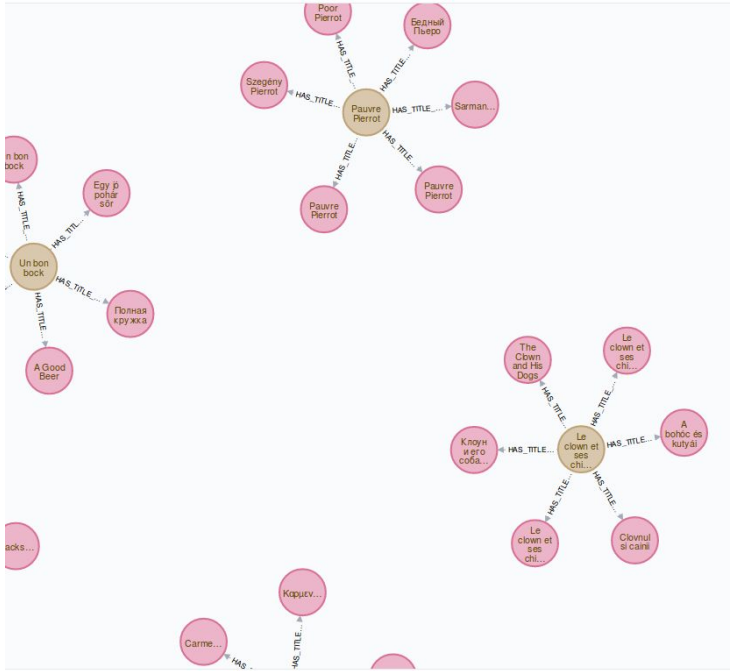

Abb.: Übersetzungen von Titeln

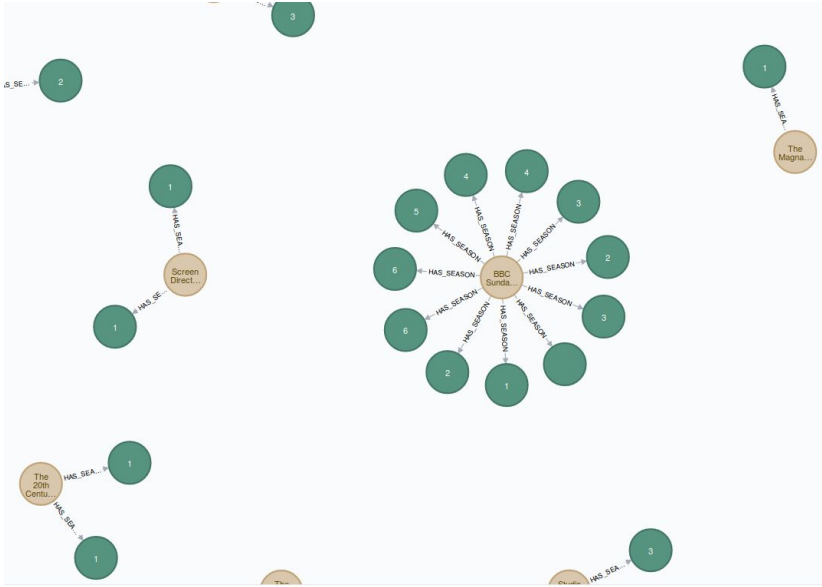

Abb.: Staffeln im Bezug auf Serien

Vor- und Nachteile

- Die Geschwindigkeit lässt auch bei größeren Datenmengen nicht nach
- Man kann sehr schnell und einfach nach/über Beziehungen suchen
- Die Beziehungen sind indexiert, weshalb die Datenbank beim Auslesen sehr schnell ist

- Für Analysen, die einen großen Bereich umfassen, ist ein Graph nicht geeignet
- Die Datenbank ist nicht für die Business Daten im Sinne eines Business Data Warehouses geeignet
- Nicht so gut skalierbar

Die Idee

Die Idee

- IMDb Datenbank
  - Relationale Datenbank
- Konvertierung in Graphdatenbank
- Herstellung von Beziehungen zwischen Filmen, Schauspielern, Regisseuren etc.
- Arbeiten mit Datensätzen, Erkenntnisgewinn

⇒ Was für einen Erkenntnisgewinn?

Was wollen wir herausfinden?

- Beziehungen zwischen Knoten nutzen
  - Beliebte Schauspielerkombinationen
  - Schauspieler-Regisseur-Paare
  - Erfolgreiche Genres von Schauspielern
- Klassische Abfragen?
  - Filme/Monat über die Zeitspanne der Daten

- Erfahrungen mit Graphdatenbanken, Vor- & Nachteile ausloten

Umsetzung

Umsetzung

- Docker Container mit neo4j auf Linux Server
- .tsv Datei der IMDb heruntergeladen
- Über MariaDB eine relationale Datenbank aus der .tsv erstellt, um dann aus der Datenbank eine .csv zu exportieren
- Der Datensatz musste erst bereinigt werden, da viele fehlerhafte Einträge vorhanden waren
- Aus der .csv mithilfe von Cypher eine Graphdatenbank erstellt

Ergebnisse

Fazit

- Es funktioniert (naja)
- Cypher ist super (also es tut halt irgendwas)

- Erkenntnisgewinn?
  - Mit Graph Datenbank gearbeitet
  - Beziehungen zwischen Daten genutzt
  - Einige komplexe Sachen nicht umgesetzt
  - IMDb Chaos genossen

Repository & Dokumentation

Das Repository mit der Dokumentation kann hier gefunden werden:
https://github.com/NerdyStuff/Other_Databases