Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Unstrukturierte Daten in einem DWH

Damit der Wert von Unternehmungen besser eingeschätzt werden kann, möchten
Analysten eine Korrelation zwischen Nachrichten und Aktienkursen herstellen,
um diese zu analysieren. Da jede Nachricht individuellen Charakter und eine
hohe Volatilität vorweist, lassen sich diese nicht in einer strukturierten Weise in einem Data Warehouse speichern.

In dieser Bachelorarbeit haben wir deshalb analysiert, wie semi- bzw. unstrukturierte Daten in einem Data Warehouse gespeichert werden können. Zur
Identifikation von geeigneten Kandidaten analysierten wir verschiedene Datenbanktechnologien. Um die Machbarkeit zu demonstrieren implementierten wir Hive, HBase und PostgreSQL als Prototyp.

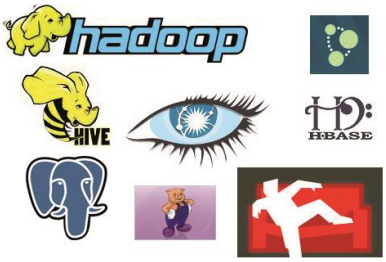
Das klassische relationale Datenbanksystem profilierte sich als bester Kandidat für kleinere Datenmengen, obschon unstrukturierte Daten nur über Umwege gespeichert werden können. Ebenso müssen keine neuen Data
Warehouse Tools eingesetzt werden. Allerdings stösst dieses System an seine
Grenzen, sobald die Datenmenge massiv zunimmt. Auch die Skalierbarkeit kann
nur schwer gewährleistet werden.

Im Gegensatz dazu hat Hive wenig Probleme mit der Skalierbarkeit und der Speicherung unstrukturierter Daten. Hive kann auch sehr grosse Datenmengen
ohne Mühe verwalten und auswerten. Die Abfragesprache von Hive ist sehr ähnlich
zu den Abfragesprachen von relationalen Datenbanksystemen. Allerdings eignet
sich Hive nur bei sehr grossen Datenmengen, da die Performance bei kleinen
Mengen verhältnismässig schlecht ist. Zudem ist Hive inkompatibel zu den
bestehenden Data Warehouse Tools.

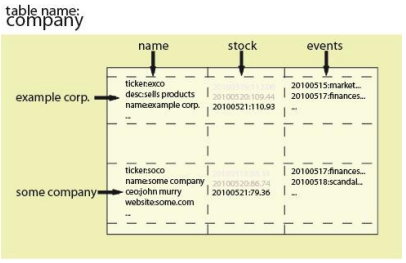
HBase ist optimal ausgelegt, um unstrukturierte Daten zu verwalten. Allerdings
ist die Auswertung der Daten sehr komplex. Aufgrund dieser Komplexität raten
wir vom Einsatz dieses Systems ab.

Diplomierende
Andreas Arnold
René Florian Poyyayil

Dozierende
Gerold Baudinot
Martin Braschler



Für diese Bachelorarbeit haben wir neben den klassischen Datenbanken auch diverse neuere Technologien evaluiert. Um die Machbarkeit zu demonstrieren, implementierten wir Hive, HBase und PostgreSQL als Prototyp.



HBase ermöglicht es, unstrukturierte Daten in ein flexibles Schema zu speichern. Einträge werden über Keys (example corp.) definiert und in Familien (z.B. name) unterteilt. Die Anzahl Attribute kann für jeden Eintrag frei definiert werden.