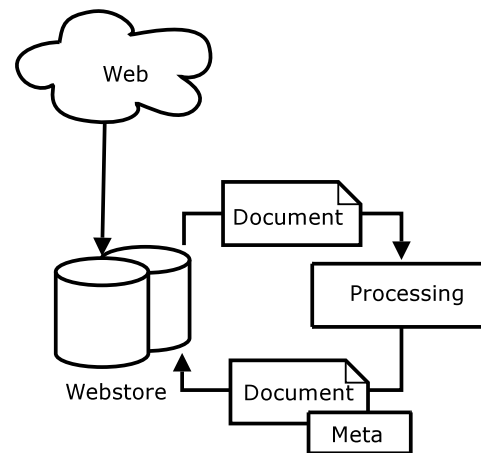


Hadoop/Hbase als Webstore

Rasmus Hahn <rassahah@neofonie.de>

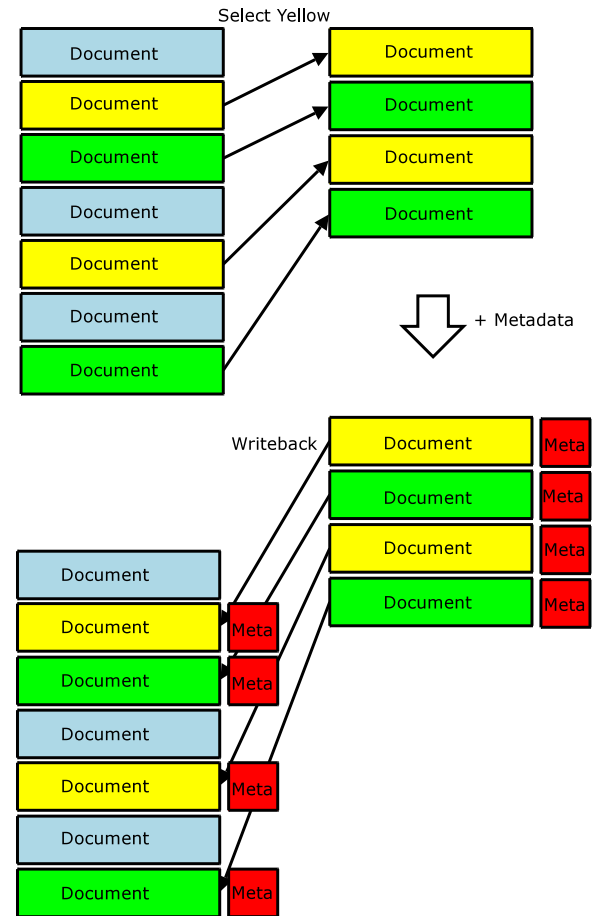
Hadoop/Hbase als Webstore: Aufgabe

- Spidern (laden) von Internetdokumenten
- Zwischenspeichern der Dokumente
- Anreichern der Dokumente mit Metadaten



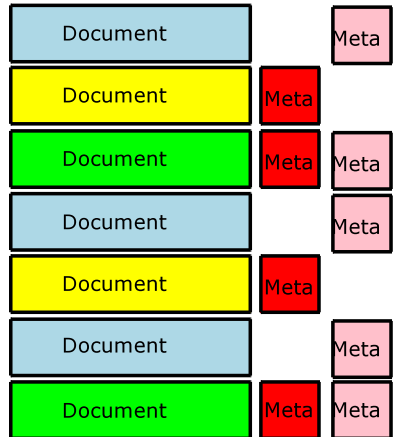
Aufgabe: Metadaten speichern

- Subset bilden
- Metadaten erzeugen
- Metadaten speichern



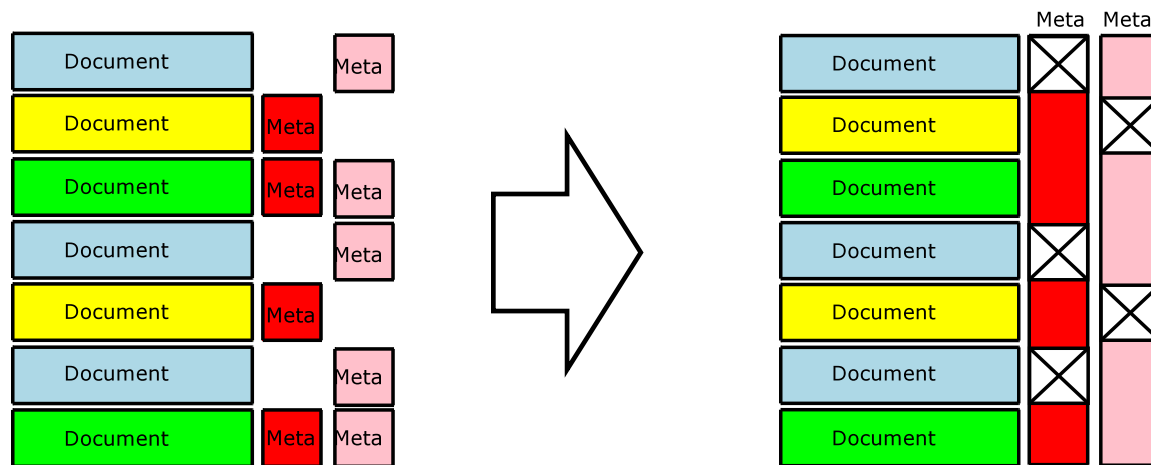
Aufgabe: Mehrere *Typen* von Metadaten

- Verschiedene Anwendungen
→ unterschiedliche Metadaten



Aufgabe: Mehrere *Typen* von Metadaten

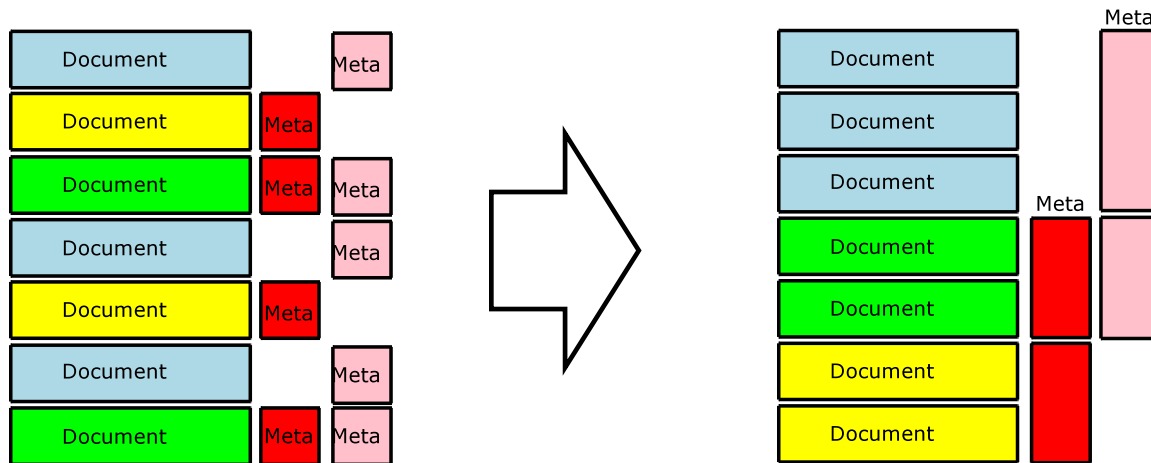
- Verschiedene Anwendungen
→ unterschiedliche Metadaten
- Jedes Dokument bekommt alle Metadaten



Problem: Speicherverbrauch und Updates

Aufgabe: Mehrere *Typen* von Metadaten

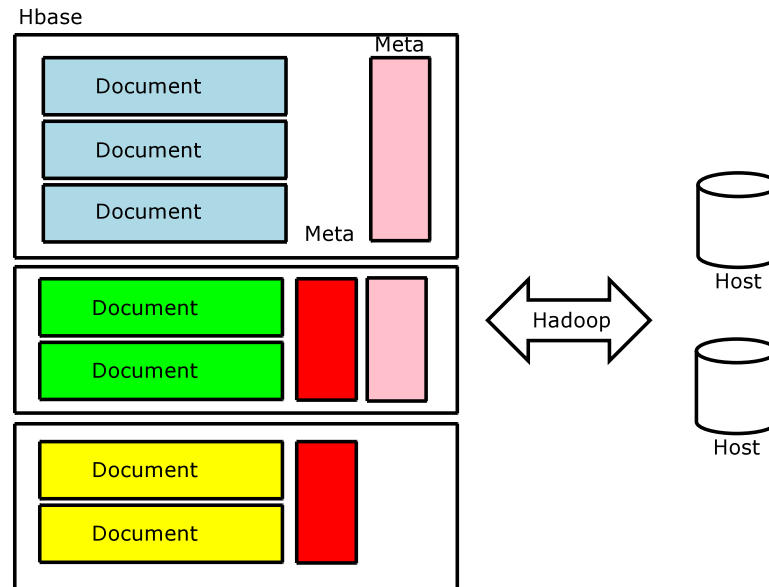
- Verschiedene Anwendungen
→ unterschiedliche Metadaten
- Verschiedene Spaltensätze von Metadaten



Komplizierte Verwaltung → *hbase*

Hbase: Vertikale Teilung der Daten

- Aufteilung der Spalten in *Regions*
- Regions werden als Datei in Hadoop gespeichert
- Zentraler *Regionserver* übernimmt Verwaltung



Probleme

- Stabilität
- Spezialisierte API
- Anpassung im Denken

Problem: Stabilität

- Verwendete Version: 0.2.0 pre-release; Abstürze tendieren zu Datenverlust
- Probleme bei der verteilten Synchronisierung; schwer aufzufinden
- Gleichmäßige Lastverteilung
- Bevorzugung des lokalen Knotens (hadoop)

Problem: Spezialisierte API

- Map-Reduce nicht universell genug
- Aufgaben müssen auf API *passen*
- Minimale Operationsmenge macht Anpassungen an Algorithmen erforderlich

Problem: Umstellung im Denken

- OO-Idee: Dokumente sind Objekte; Metadaten werden durch Messages erzeugt.
- hbase: Relationaler Aufsatz auf einfache Daten; Speicherung o.k.
- Map-Reduce: Funktionale Herkunft, Funktionen von Daten getrennt