

Zitierungsanalyse mit Kylin

Big OLAP: Data-Warehouse

Sebastian Lange, Simon Hüning, Klemens Schölhorn

04.03.2016

Einleitung und Aufgabenstellung

Data-Warehouse (DWH)

- Analyseabfragen
- Star-Schema
 - **Faktentabelle** Messwerten, Kennzahlen, ...
 - **Dimensionstabellen** Beschreibende Daten (Autorenname, Buchtitel, ..)
- OLAP-Würfel

Aufgabenstellung

- Analyse des Ausgangsdatensatzes
 - DBLP - Sammlung wissenschaftlicher Publikationen
- Importierung in das HDFS
- Transformation des Datensatzes in das Star-Schema
- Erzeugung der OLAP-Würfel
- Vergleich Kylin-Anfragen/HQL-Anfragen

└ Einleitung und Aufgabenstellung

Einleitung und Aufgabenstellung

Data-Warehouse (DWH)

- Analyseabfragen
- Star-Schema
 - **Fakientabelle**: Messwerten, Kennzahlen, ...
 - **Dimensionstabellen**: Beschreibende Daten (Autorenname, Buchtitel, ...)

• OLAP-Würfel

Aufgabenstellung

- Analyse des Ausgangsdatensatzes
 - DBLP - Sammlung wissenschaftlicher Publikationen
- Importierung in das HDFS
- Transformation des Datensatzes in das Star-Schema
- Erzeugung der OLAP-Würfel
- Vergleich Kylin-Anfragen/HQL-Anfragen

- Ein Data Warehouse (DW oder DWH) ist eine für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, in der Regel heterogenen Quellen zusammenführt und verdichtet.
- Faktentabelle - Foreignkeys + Fakten
- Dimensionstabelle - Beschreibende Daten, Name, Alter, etc...
- Innerhalb wissenschaftlicher Arbeiten werden andere Arbeiten zitiert
- Anzahl der Zitierungen charakterisiert wissenschaftlichen Einfluss

Verwendete Technologien

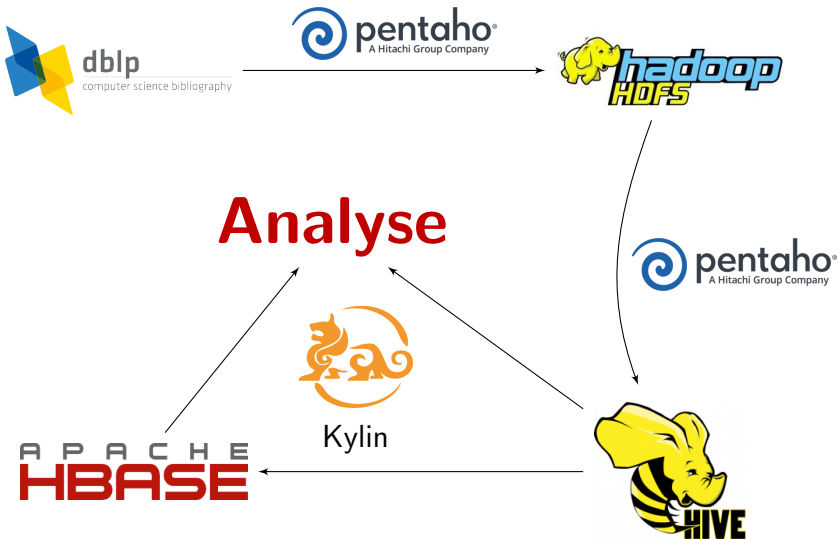
- Hadoop 2.7.1
- HBase 0.98.15 (Hadoop 2.2 Bibliotheken)
- Hive 0.14.0
- Kylin 1.2
- Flink 0.10.1 (Hadoop 2.7 Bibliotheken)
- MySQL 5.5.47 (Hive-Metastore)

└ Verwendete Technologien

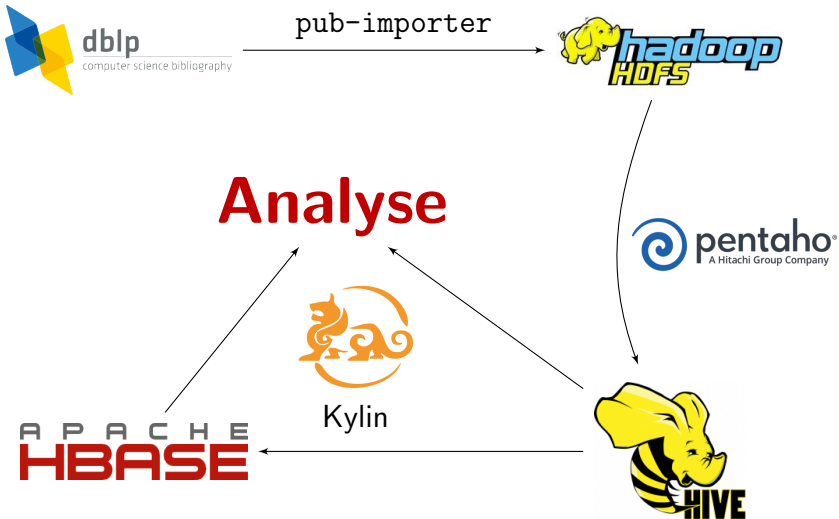
- Hadoop 2.7.1
- HBase 0.98.15 (Hadoop 2.2 Bibliotheken)
- Hive 0.14.0
- Kylin 1.2
- Flink 0.10.1 (Hadoop 2.7 Bibliotheken)
- MySQL 5.5.47 (Hive-Metastore)

- Hadoop/HDFS - MapReduce-Framework/verteiltes Dateisystem - single node setup
- HBase - Key/Value-Store
- Hive - DWH-Erweiterung für Hadoop mit eigener Abfragesprache (HQL)
- Kylin - OLAP-Engine
- Flink - Hadoop-ETL-Framework

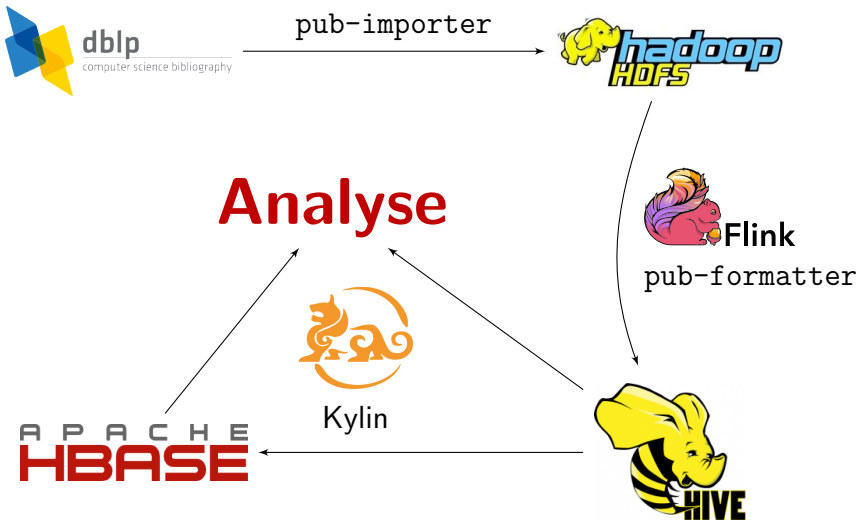
Datenfluss



Datenfluss




Datenfluss



Import: pub-importer

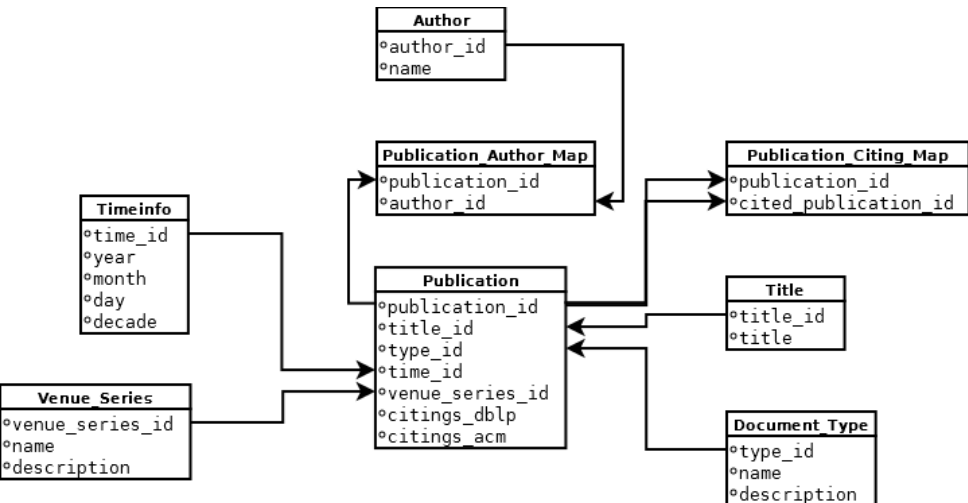
```
<article mdate="2012-09-16" key="journals/tkde/FanS93a">
  <author>Jang-Jong Fan</author>
  <author>Keh-Yih Su</author>
  <title>Corrections to "An Efficient Algorithm
    for Matching Multiple Patterns".</title>
  <year>1993</year>
  <volume>5</volume>
  <journal>IEEE Trans. Knowl. Data Eng.</journal>
  <number>5</number>
  <url>db/journals/tkde/tkde5.html#FanS93a</url>
  <cite>journals/tkde/FanS93</cite>
</article>
```



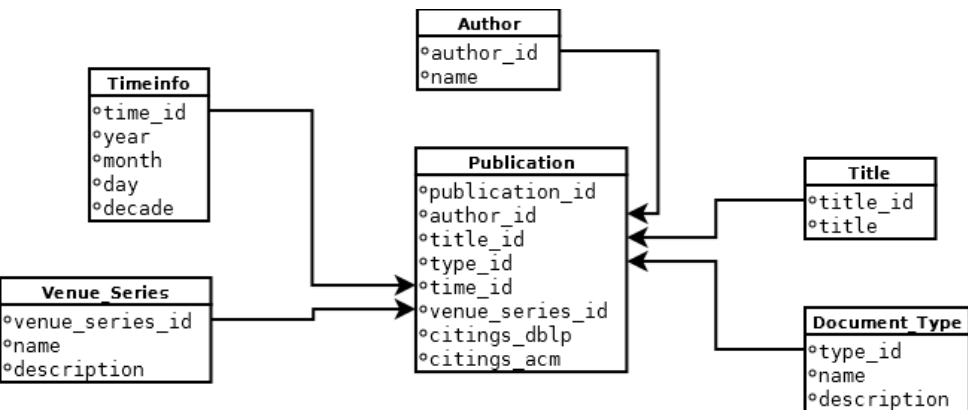
```
curl http://dblp.uni-trier.de/xml/dblp.xml.gz | gunzip
| java -jar pub-importer.jar dblp /path/in/hdfs
```

7802,article,journals/tkde/FanS93a,Corrections to "An Efficient Algorithm for Matching Multiple Patterns".,1993,IEEE Trans. Knowl. Data Eng.#5#5,Jang-Jong Fan|Keh-Yih Su,journals/tkde/FanS93

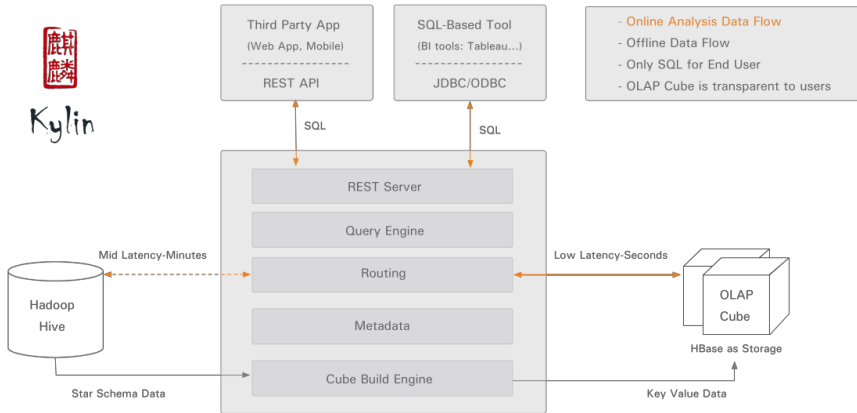
Transformation: pub-formatter



Transformation: pub-formatter



Analyse: Kylin



└ Analyse: Kylin

- MOLAP für Hadoop
- Nutzung von HBASE für Cube Generierung
- SQL Engine (Apache Calcite)
- Adhoc-Analyse

Analyse: Kylin



Probleme und Einschränkungen

- Pentaho Data Integration
 - StAX Parser schlecht integriert
 - JOINS komplett in Zwischenspeicher
 - Performance Probleme bei kompletter Datenmenge (12 GB RAM nicht genug)
 - Umständliche Konfiguration für Nutzung auf Cluster
- Flink
 - Kein standardkonformer CSV-Reader
- Hive
 - Inkorrekte Speicherberechnungen (alte Version)
- Kylin
 - Keine VIEW als Lookup-Table verwendbar
 - SQL Engine hat Probleme mit DISTINCT

Ergebnisse

	HIVE	KYLIN	Ersparnis
# Pub/Autor	177,34 s	51,33 s	126,01 s
# Pub/Autor/2015	76,88 s	7,55 s	69,33 s
max Zit/Autor/Pub	120,73 s	15,37 s	105,36 s
max Zit/Autor	317,96 s	31,80 s	286,16 s
# Pub/Rahm	70,70 s	0,07 s	70,63 s
# Pub/Rahm/2015	59,76 s	3,71 s	56,05 s
max Zit/Pub/Rahm	184,85 s	0,32 s	184,53 s
sum Zit/Rahm	58,63 s	4,79 s	53,84 s
# Pub/Groß	94,35 s	0,13 s	94,22 s
# Pub/Groß/2015	79,25 s	0,08 s	79,17 s
Mittelwert	124,05 s	11,52 s	112,53 s

Ergebnisse

