

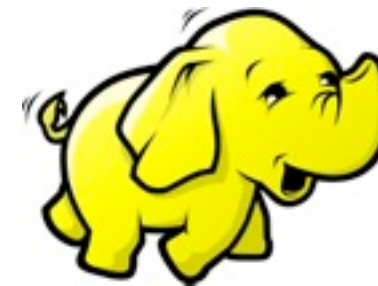


Fusion der Welten: Hadoop als DWH-Backend bei Pro7

Dr. Kathrin Spreyer
Big Data Engineer

Open Source Business Intelligence
Karlsruhe, 20.02.2014

1. **Relationales DWH**
2. **Bewirtschaftung mit PDI**
3. **Unter der Haube: Hadoop**
4. **Datenimport mit Flume**
5. **Hive im Fachbereich**

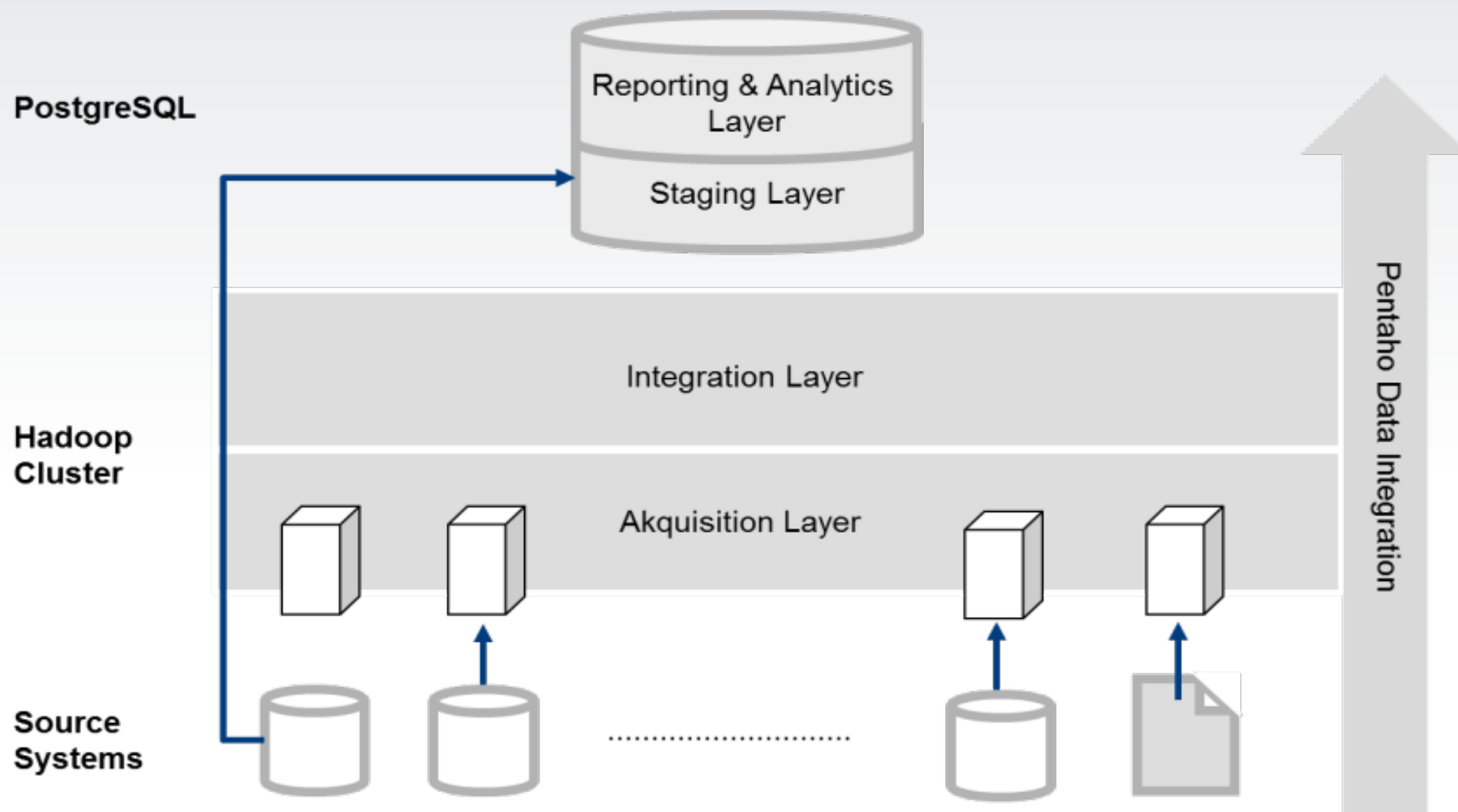


Pentaho MapReduce



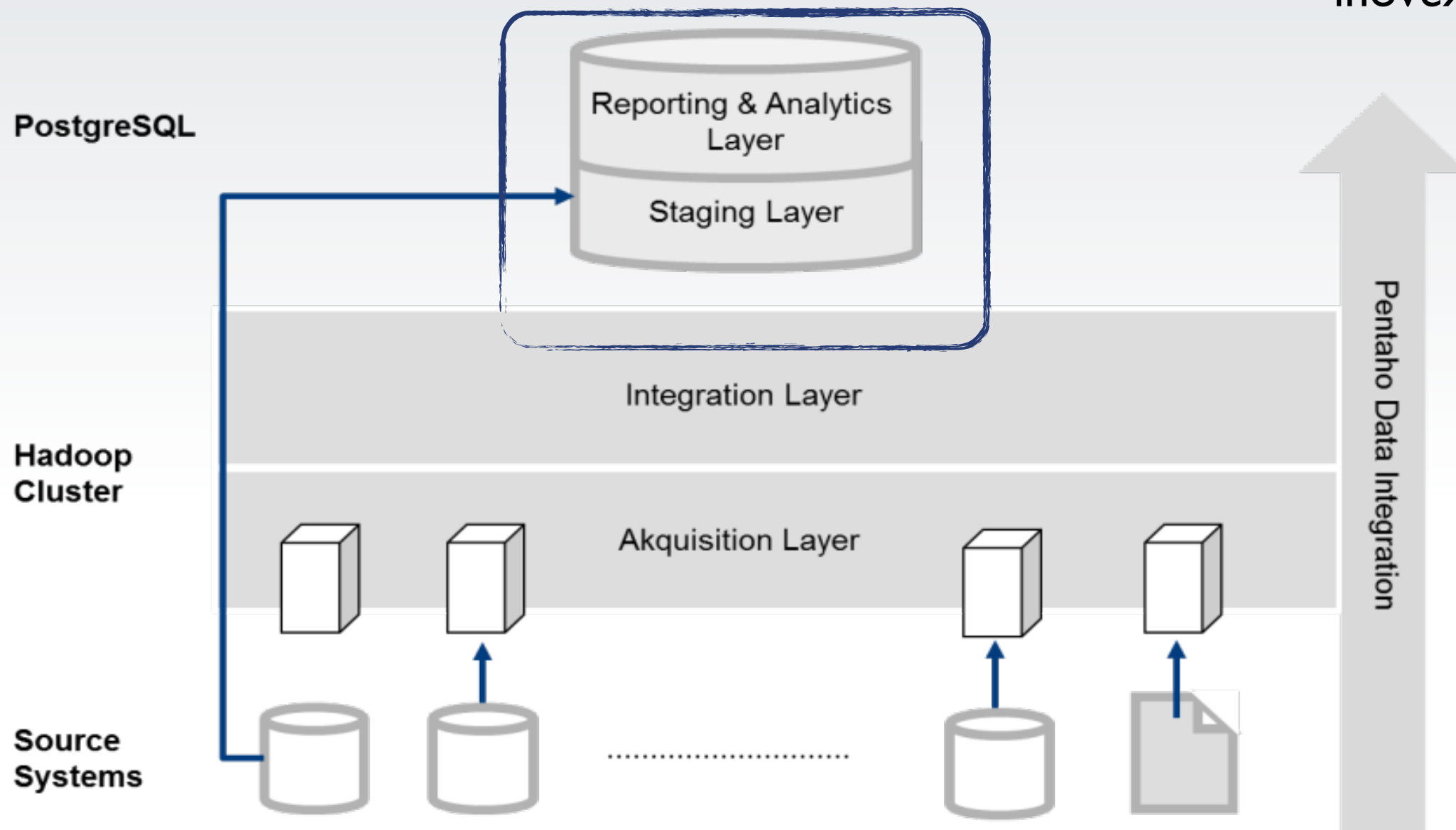
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

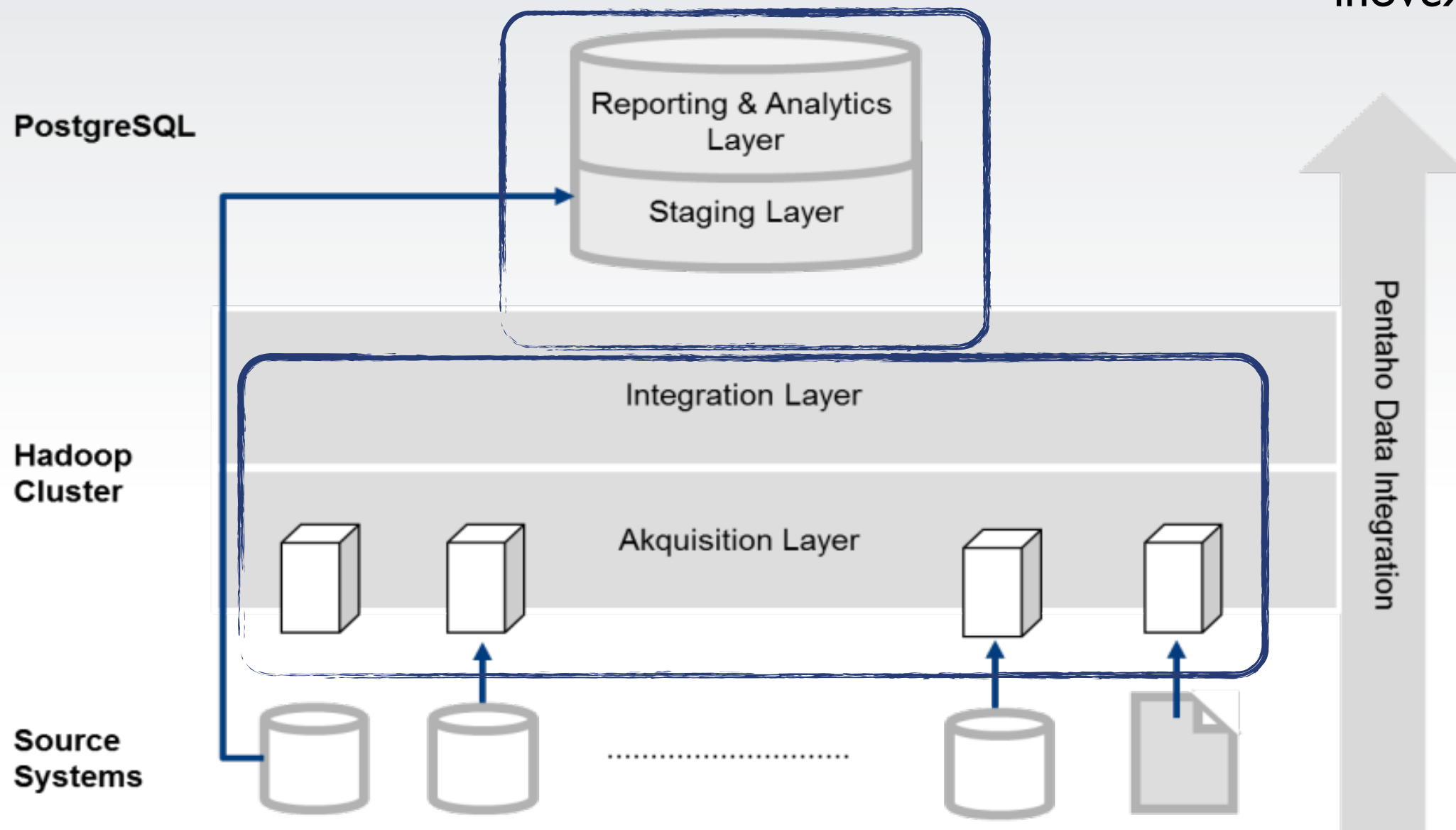
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

Relationales DWH

1. DWH zur Integration von **Reichweiten-, Vermarktungserlös- und Transaktionsdaten**



ProSiebenSat.1 **Digital** Wesentlicher Treiber der Digitalstrategie

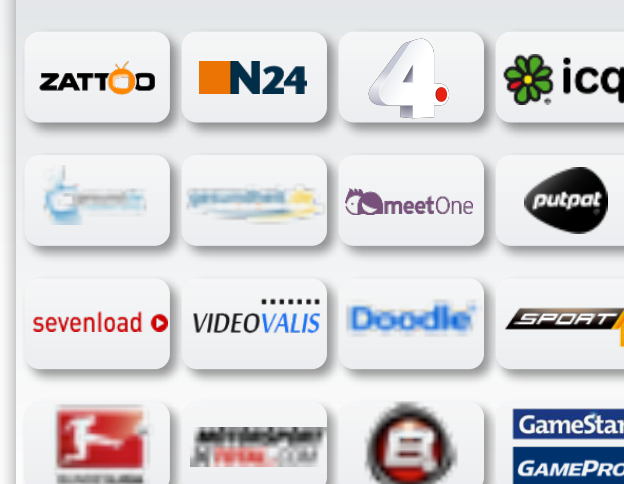
Online TV Channel

MyVideo

ProSiebenSat.1 Network



Externe Mandanten



1. DWH zur Integration von **Reichweiten-, Vermarktungserlös- und Transaktionsdaten**



ProSiebenSat.1 **Digital** Wesentlicher Treiber der Digitalstrategie

35 Online Platforms (incl. external mandates)

26m Ø Unique User per month

143m Video Views per month

max(dome) axel springer media impact sevenload N24

4 AMPYA autoplenum.de wer weiss was

MOTORSPORT TOTAL.COM FORMEL1.de SPORT Doodle

MOTORVISION.DE wetter.com gesundheit.de MOTOR TALK.de

viewster gesund.tv Lokalstunde ran

twitch Twitch 7

BUNDESLIGA SIXX MyVideo BIGPOINT

VIDEOMARKT icq putpat BEIT TV

1. **MyVideo Apache Logs**

Metriken:

VV, VT, Ads, Player-Fehler, ..

2. **Webtrekk**

VV, VT, Ads

3. **DfP (Adserver)**

Ads

4. Google Analytics

VV, PI, Visits, Visitors

5. Soziale Netzwerke

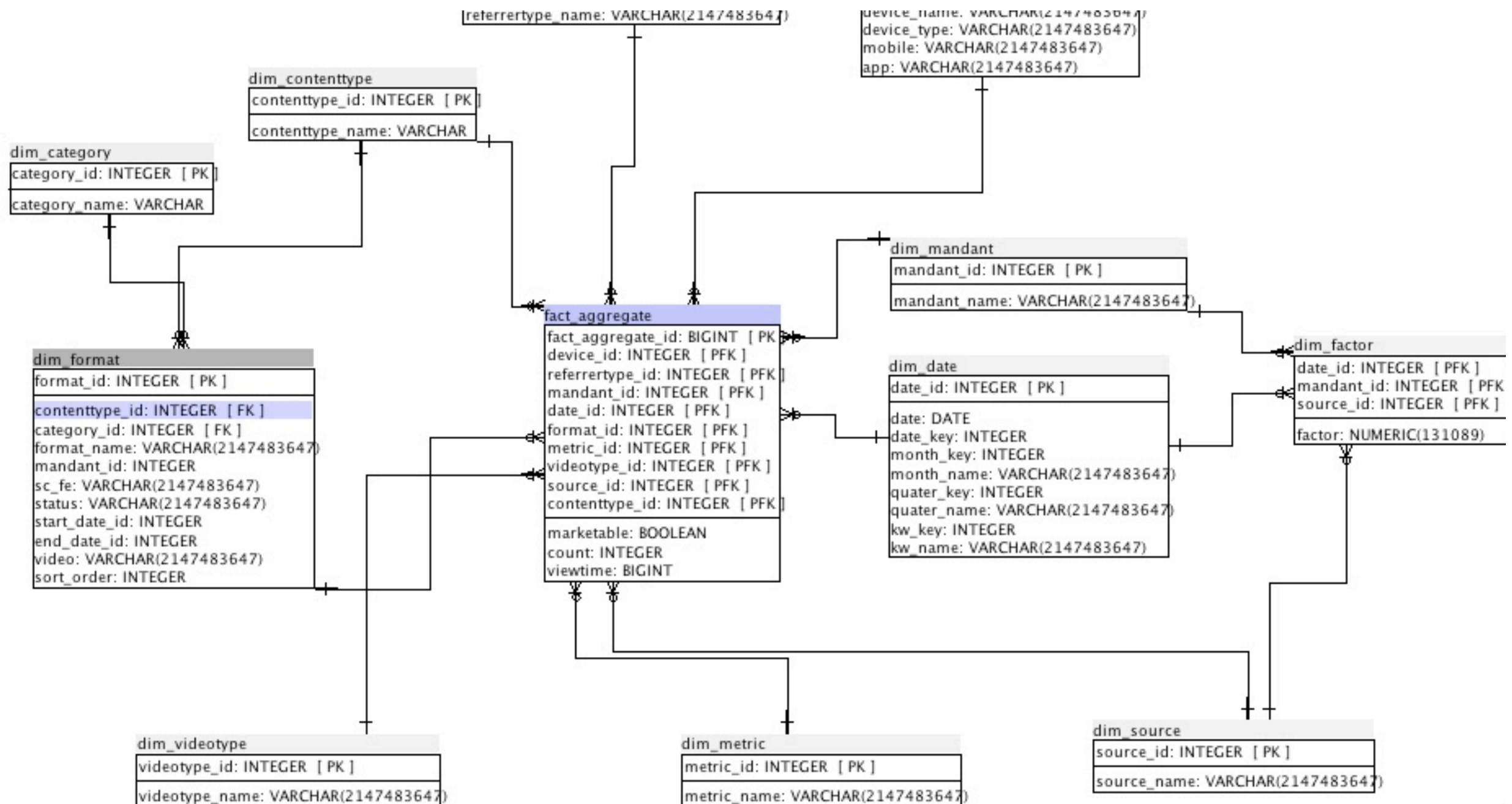
Posts, Comments, Likes

6. Conviva

Bitrate, Video startup time, ..

7.

1. Relationales Datenmodell (Sternschema)



1. Momentan viel Excel
2. Im Aufbau: Reporting mit **Business Objects**
3. Empfänger: BI, Controlling, Management, ...



DWH-Bewirtschaftung

1. ETL-Tool **Pentaho Data Integration** (PDI)

1. Datenakquise
2. Validierung / Reinigung / Fehlerbehandlung
3. Aggregation
4. Export ins DWH

1. **Konnektoren** zu gängigen Daten-/Speicherformaten

1. **Dateien:** CSV, JSON, XML, ...

2. **RDBMS:** 40+ connection types

3. **Big data:** HDFS, Cassandra, HBase, MongoDB, ...

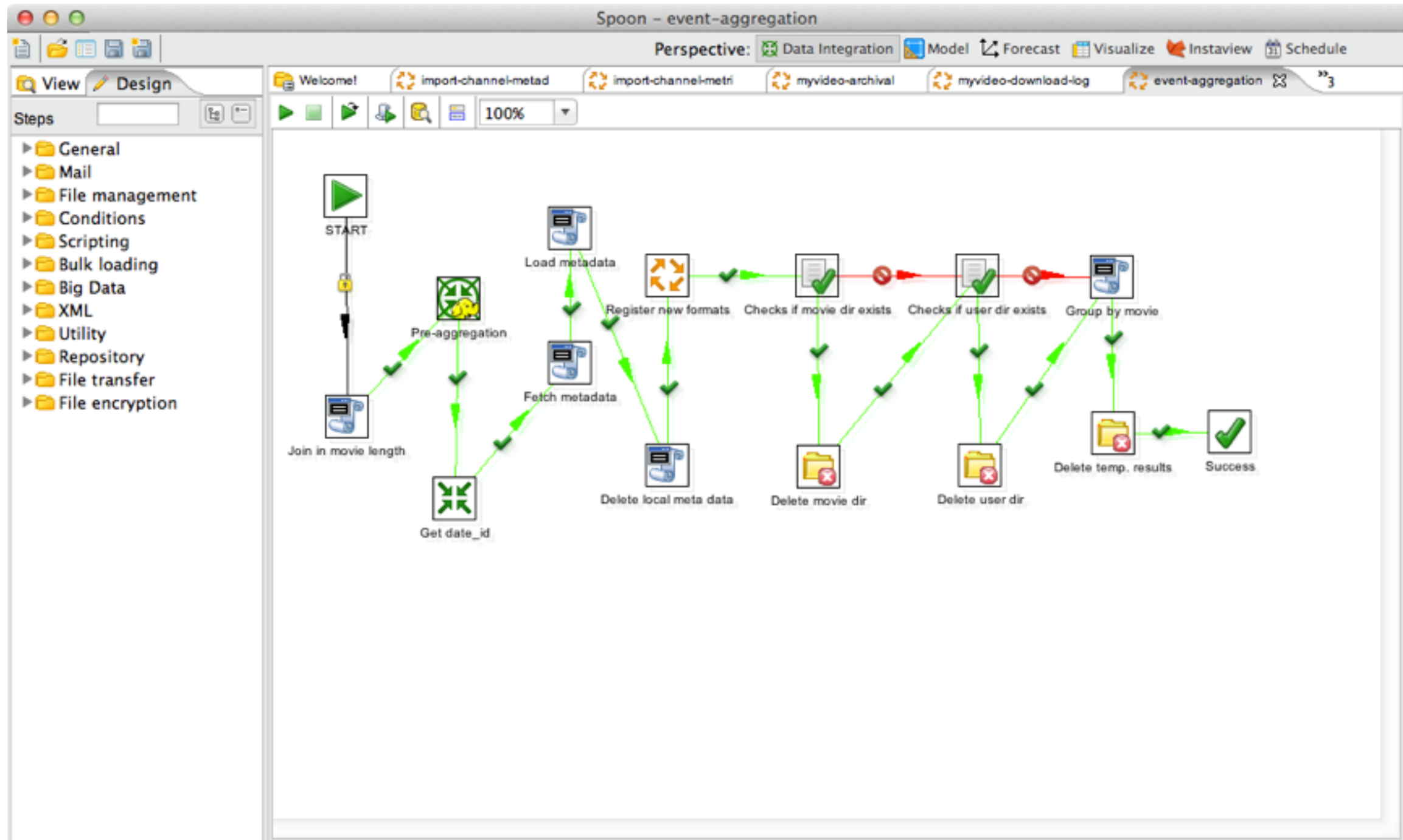
1. **Konnektoren** zu gängigen Daten-/Speicherformaten

1. **Dateien:** CSV, JSON, XML, ...

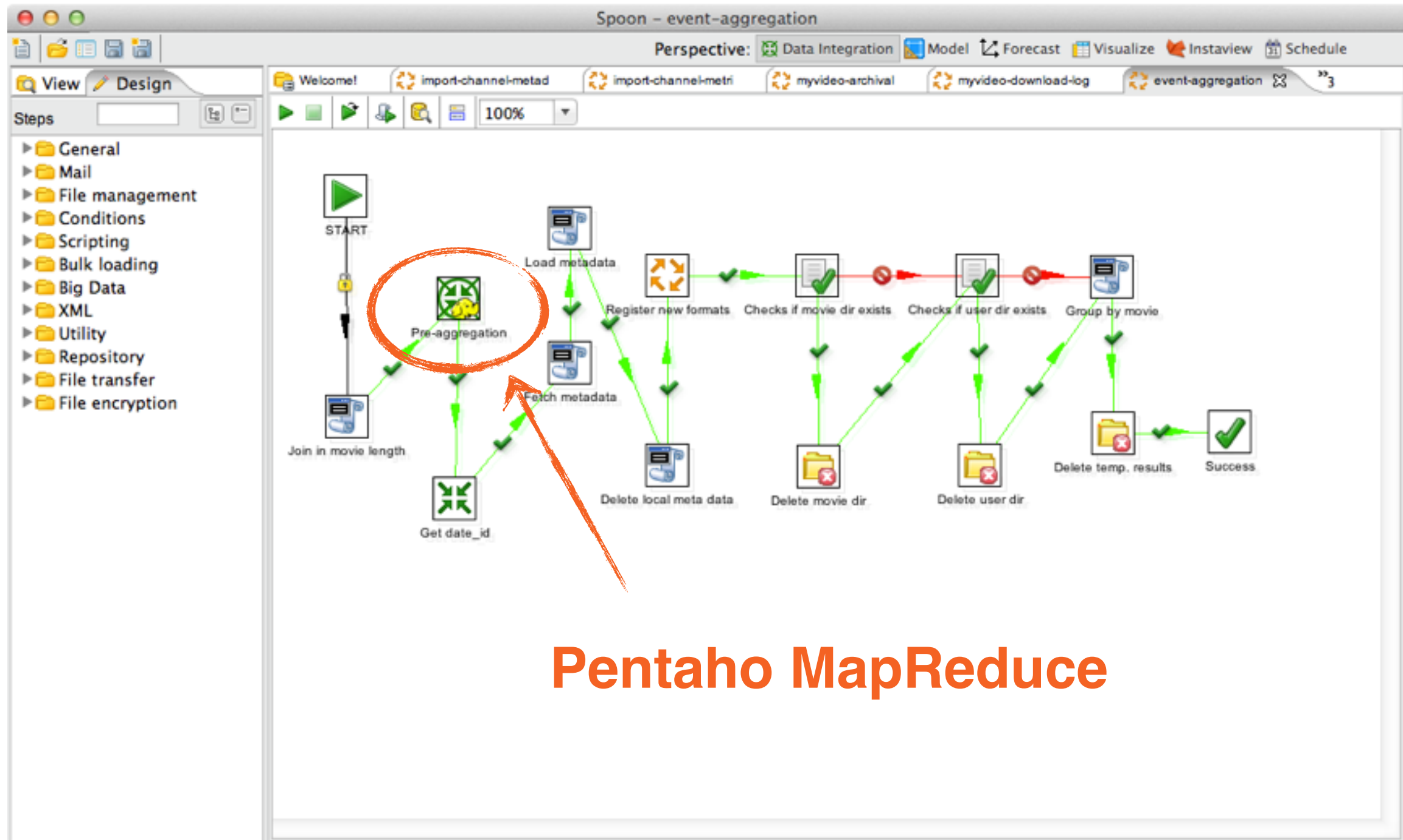
2. **RDBMS:** 40+ connection types

3. **Big data:** HDFS, Cassandra, HBase, MongoDB, ...

Pentaho Data Integration

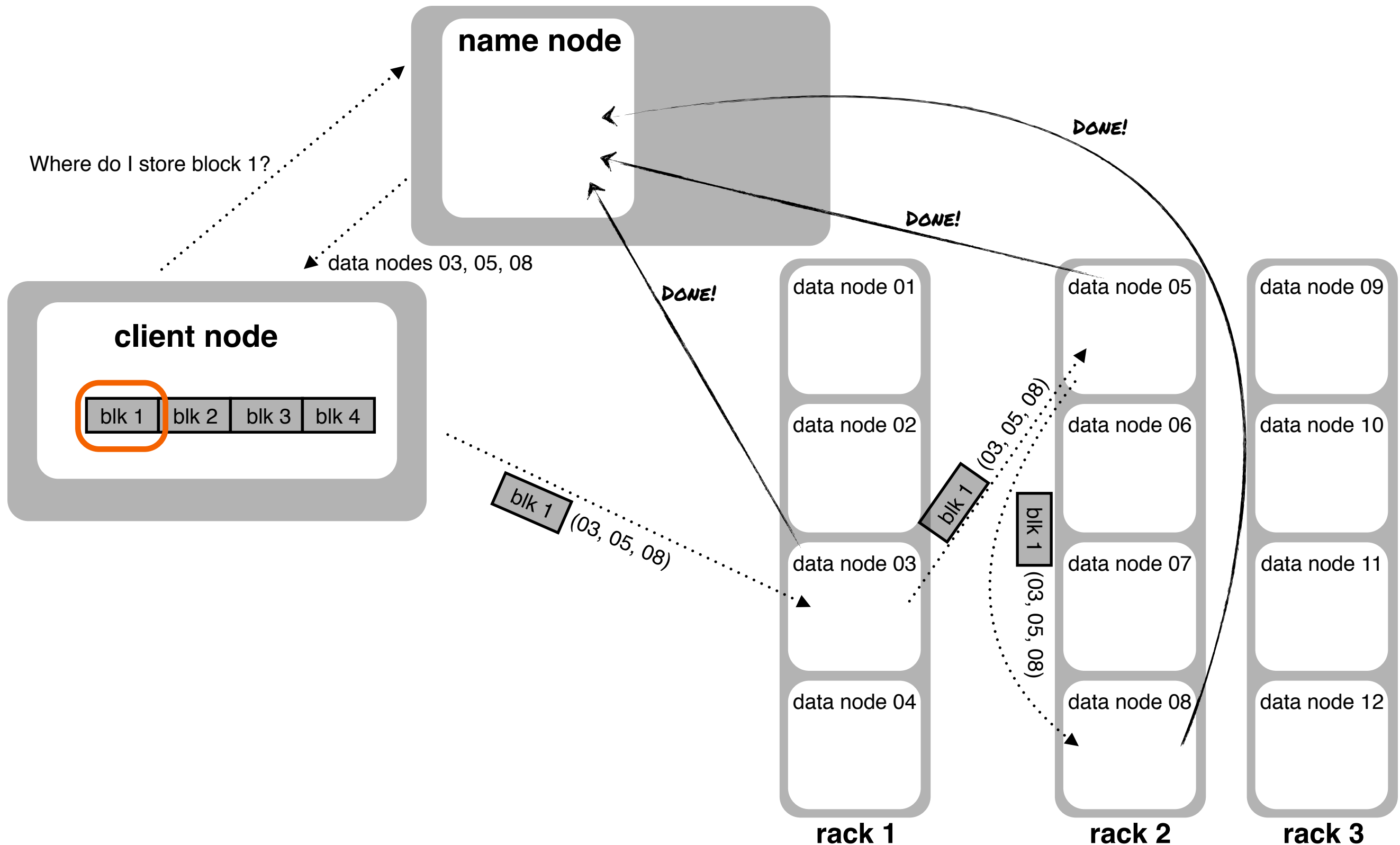


Pentaho Data Integration

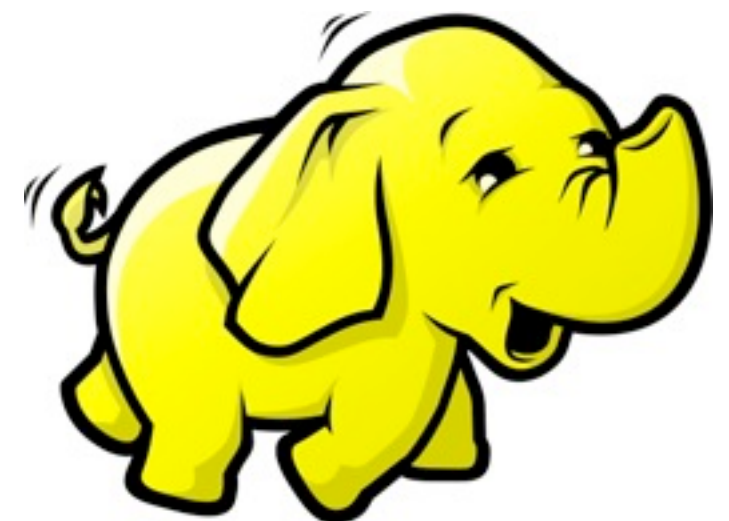


Pentaho MapReduce

Hadoop-Backend



1. **Aggregation** der Rohdaten
2. Transparente **Parallelisierung**
3. Horizontal **skalierbar**



Java: Mapper und Reducer

```
public class WebtrekkEventMapper
    extends Mapper<Text, Text, Text, IntWritable> {

    @Override
    protected void map( Text key, Text value, Context context )
        throws IOException, InterruptedException {
        // key contains entire record
        String[] fields = key.toString().split( ";" );
        // extract relevant information
        String eventname = fields[12];
        // emit output key and count
        context.write( new Text( eventname ),
            new IntWritable( 1 ));
    }
}

public class IntSumReducer
    extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    protected void reduce( Text key, Iterable<IntWritable> values,
        Context context )
        throws IOException, InterruptedException {
        int sum = 0;
        for ( IntWritable partialCount : values ) {
            sum += partialCount.get();
        }
        context.write( key, new IntWritable( sum ) );
    }
}
```

Java: Mapper und Reducer

Mapper

(+Driver)

Reducer

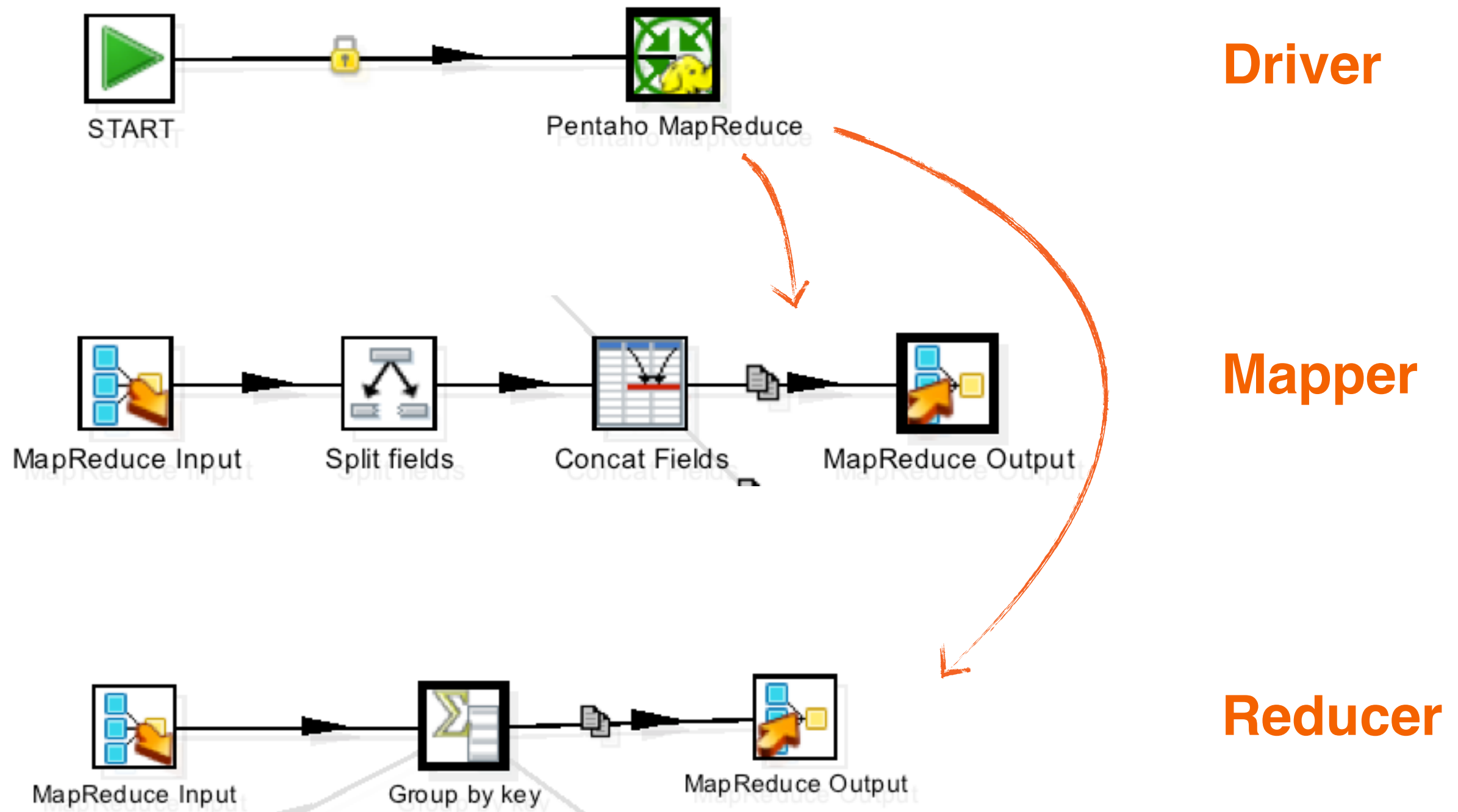
```
public class WebtrekkEventManager
    extends Mapper<Text, Text, Text, IntWritable> {

    @Override
    protected void map( Text key, Text value, Context context )
        throws IOException, InterruptedException {
        // key contains entire record
        String[] fields = key.toString().split( ";" );
        // extract relevant information
        String eventname = fields[12];
        // emit output key and count
        context.write( new Text( eventname ),
            new IntWritable( 1 ) );
    }
}
```

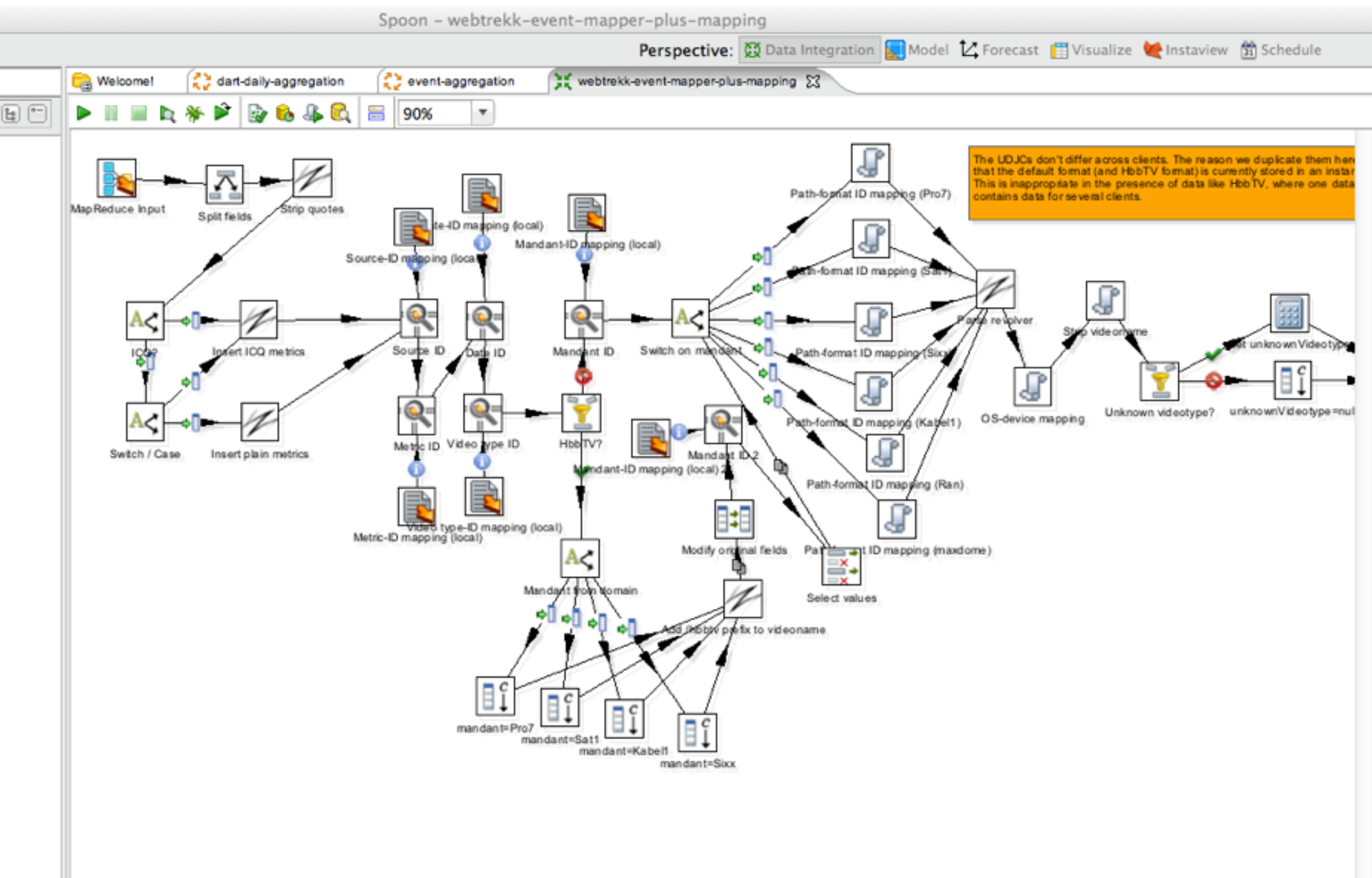
```
public class IntSumReducer
    extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
    protected void reduce( Text key, Iterable<IntWritable> values,
        Context context )
        throws IOException, InterruptedException {
        int sum = 0;
        for ( IntWritable partialCount : values ) {
            sum += partialCount.get();
        }
        context.write( key, new IntWritable( sum ) );
    }
}
```

PDI: Main, Mapper und Reducer



PDI: Mapper-Transformation



1. Unterstützt alle **PDI-Schritte**
2. **Weniger effizient** als Java MR
3. (Fast) **keine Java-Kenntnisse** nötig



1. **Native** Hadoop-API
2. Unterstützung **aller** Hadoop-Features
3. Mehr Kontrolle, **Optimierungsmöglichkeiten**



1. **Abstraktion** über MapReduce
2. **Übersetzung** in MapReduce-Jobs
3. **Nicht immer effizient**
4. Praktisch für (Equi-) **Joins**



Datenimport

1. Import via **FTP**-Server:
Daten erst nach 24 Stunden im DWH
2. **Apache Flume**: kontinuierlicher Datenstrom
3. Automatische “Einsortierung” in HDFS-Verzeichnisse
z.B. nach **Zeitstempel**



1. **Simulation** v. Datenströmen: spooling directory
2. Sonst TCP (netcat)
3. Bisher **Single-Tier-Architektur**



1. via **Flume**: jährlich 20 TB (netto)
2. via **FTP**: jährlich 21.5 TB (netto)
3. **Archivierung**: jährlich 500 GB (netto)
4. **insgesamt**: 40 TB (120 TB) pro Jahr

1. **dev:** 4 DN + NN + SN + DB + Admin
2. **prod:** 6 DN + NN + SN + DB + 2 Admin
3. **HDFS-Kapazität:** 185 TB (brutto)
4. **skalierbar**

1. **dev:** 4 DN + NN + SN + DB + Admin
2. **prod:** 6 DN + NN + SN + DB + 2 Admin
3. **HDFS-Kapazität:** 185 TB (brutto)
4. **skalierbar**

Hive im Fachbereich

1. Bekanntes Interface: **SQL**
2. Teilweise Nutzung zur **Aggregation** im Backend
3. Hauptsächlich Eruierung **neuer Metriken**
4. Untersuchung von **Daten-"Anomalien"**



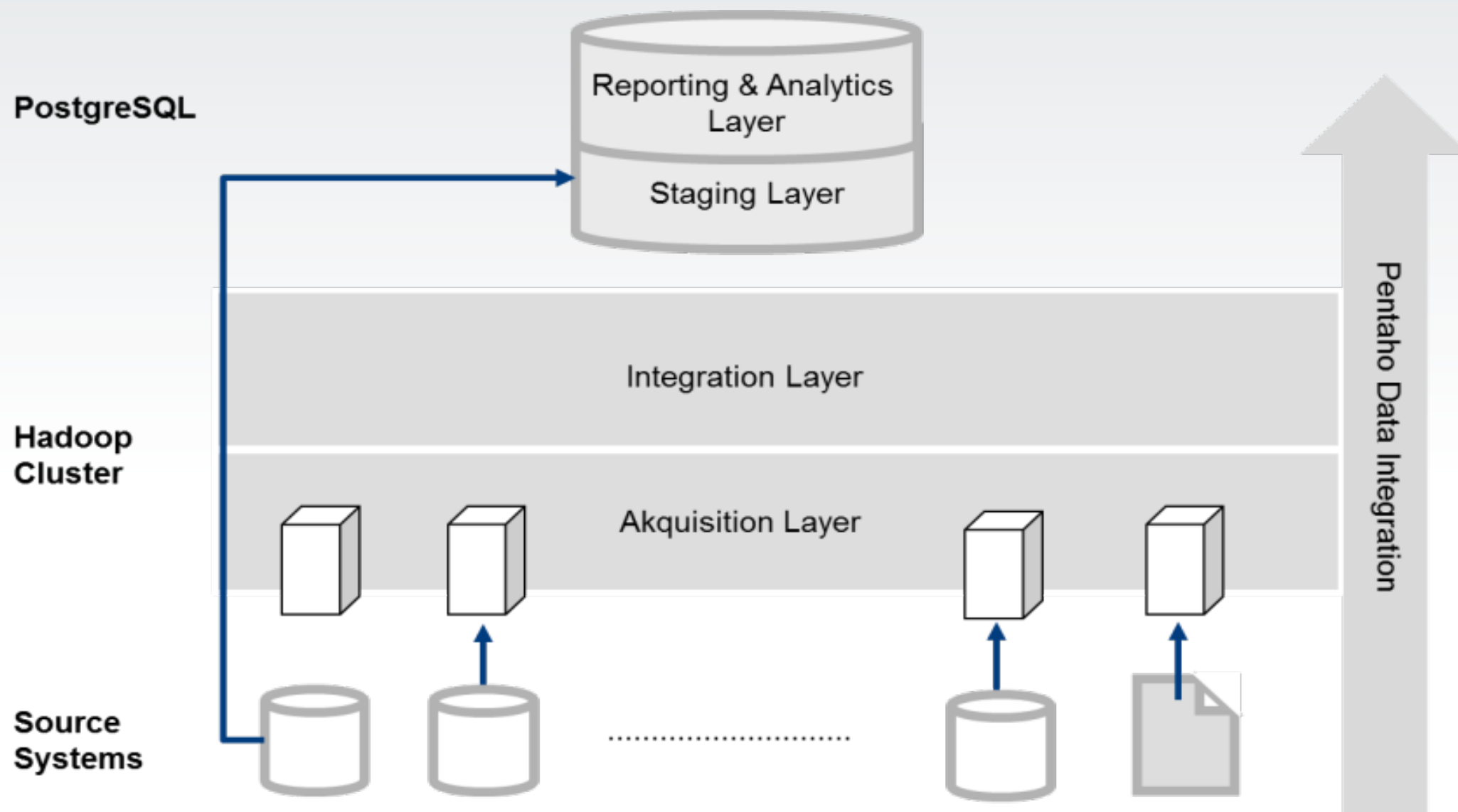
```
SELECT mandant, day, count(*)  
FROM webtrekk_cust_para_click_2  
WHERE mandant = 'sat1'  
      AND day BETWEEN '2013-01-01' AND '2013-01-31'  
GROUP BY sid, request_id, times, day, mandant  
HAVING count(*) >1
```



Hive-Zugriff auf Rohdaten

Lösungsansatz

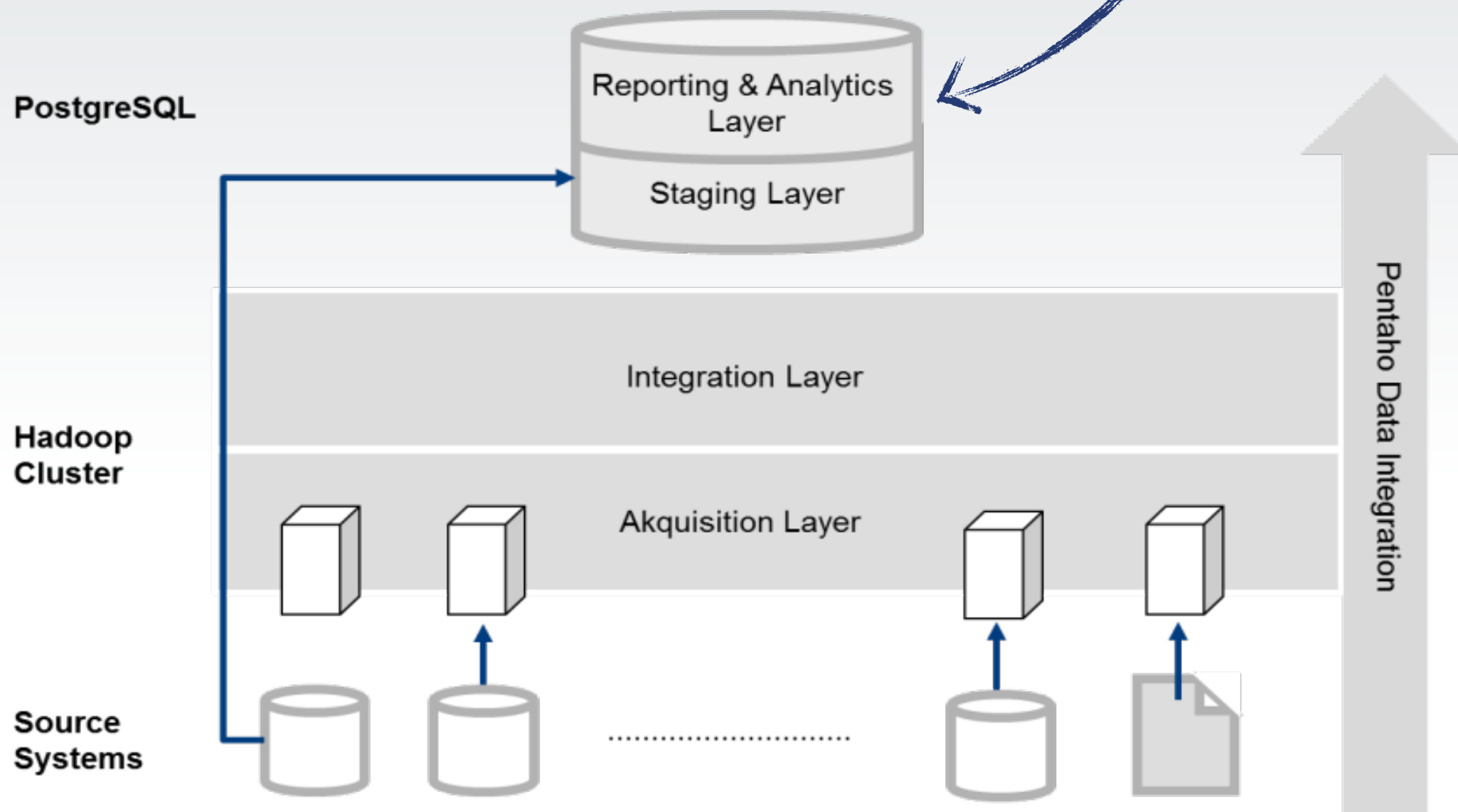
Hybrides System aus relationaler Datenbank und Hadoop Cluster



Hive-Zugriff auf Rohdaten

Lösungsansatz

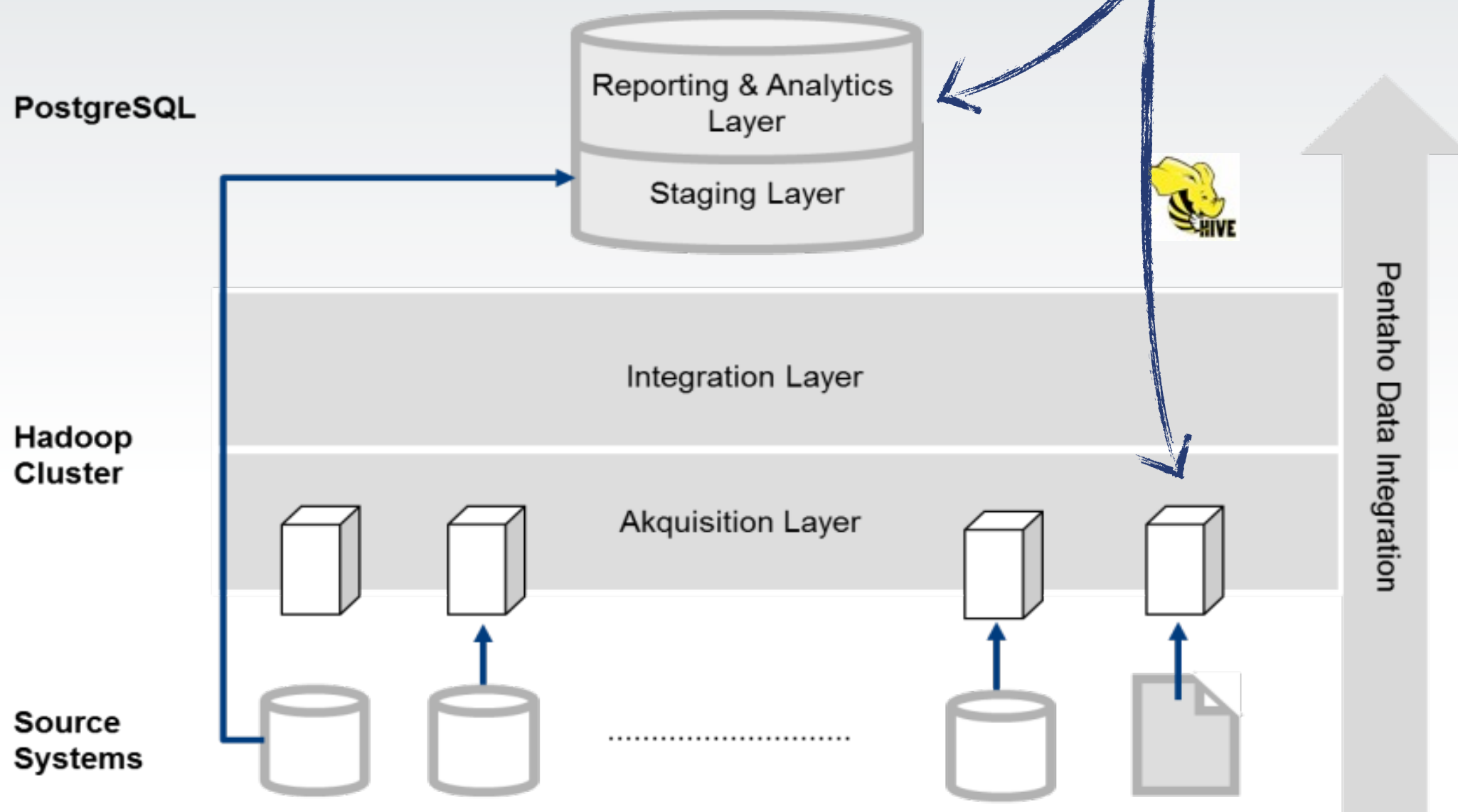
Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

Hive-Zugriff auf Rohdaten

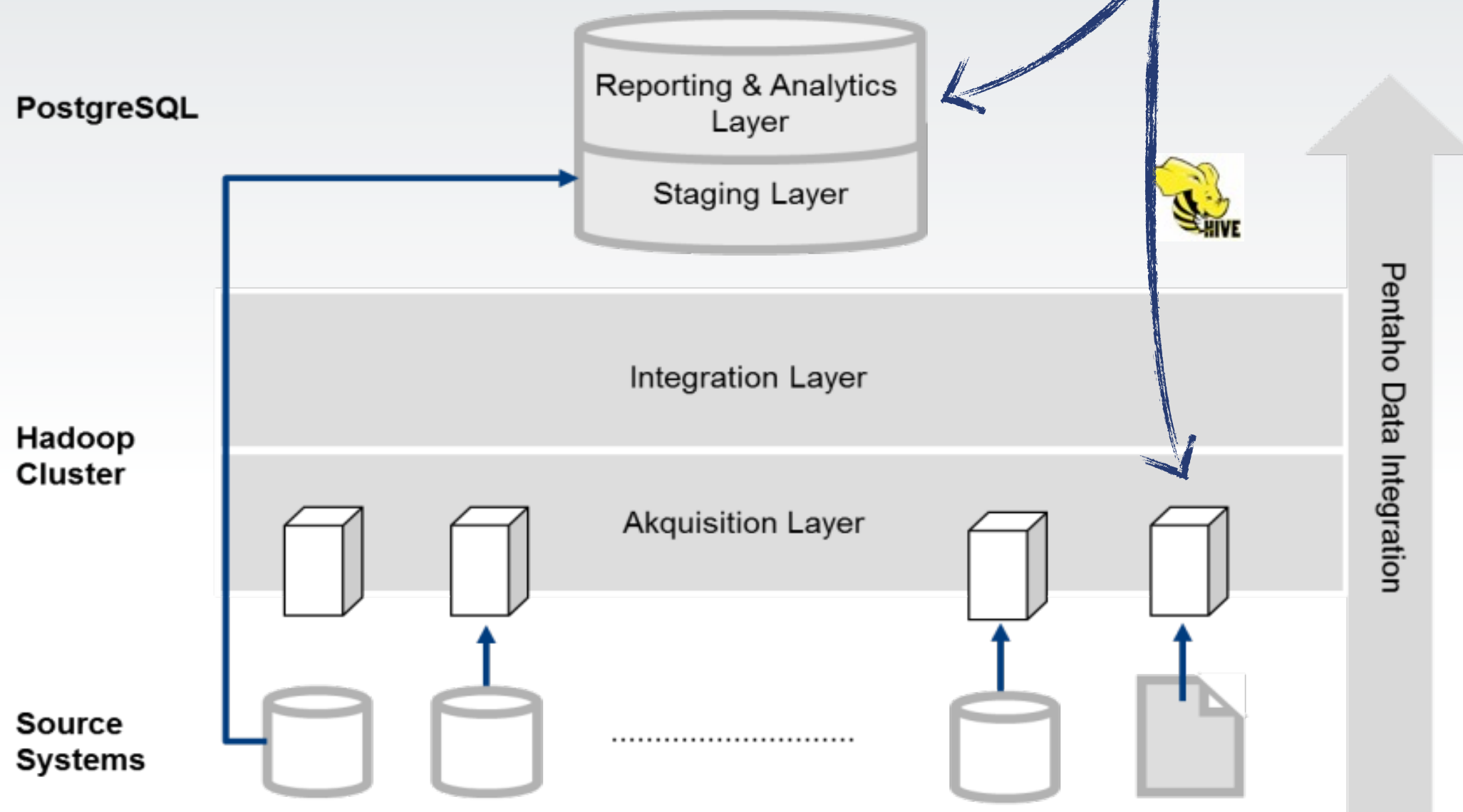
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster



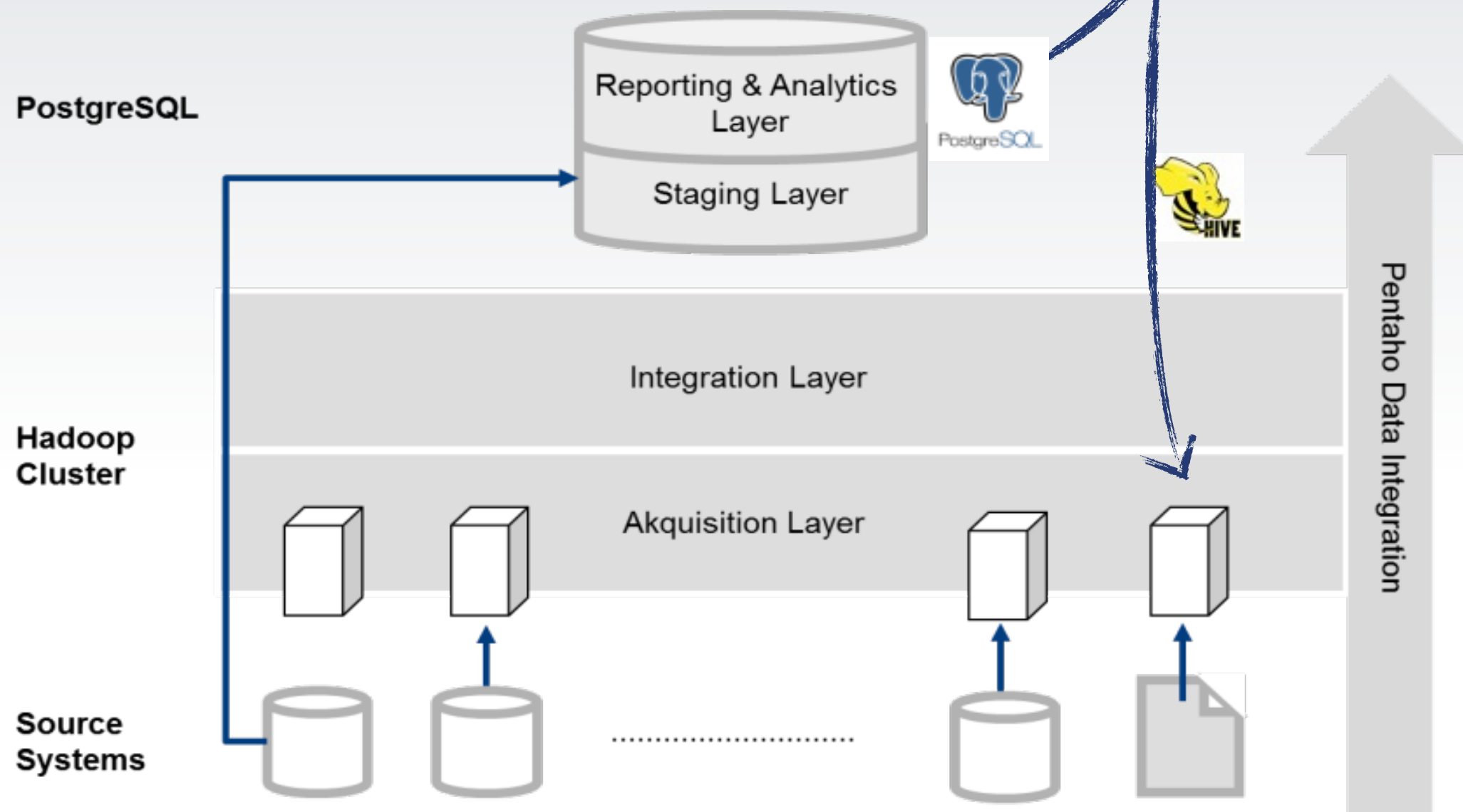
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

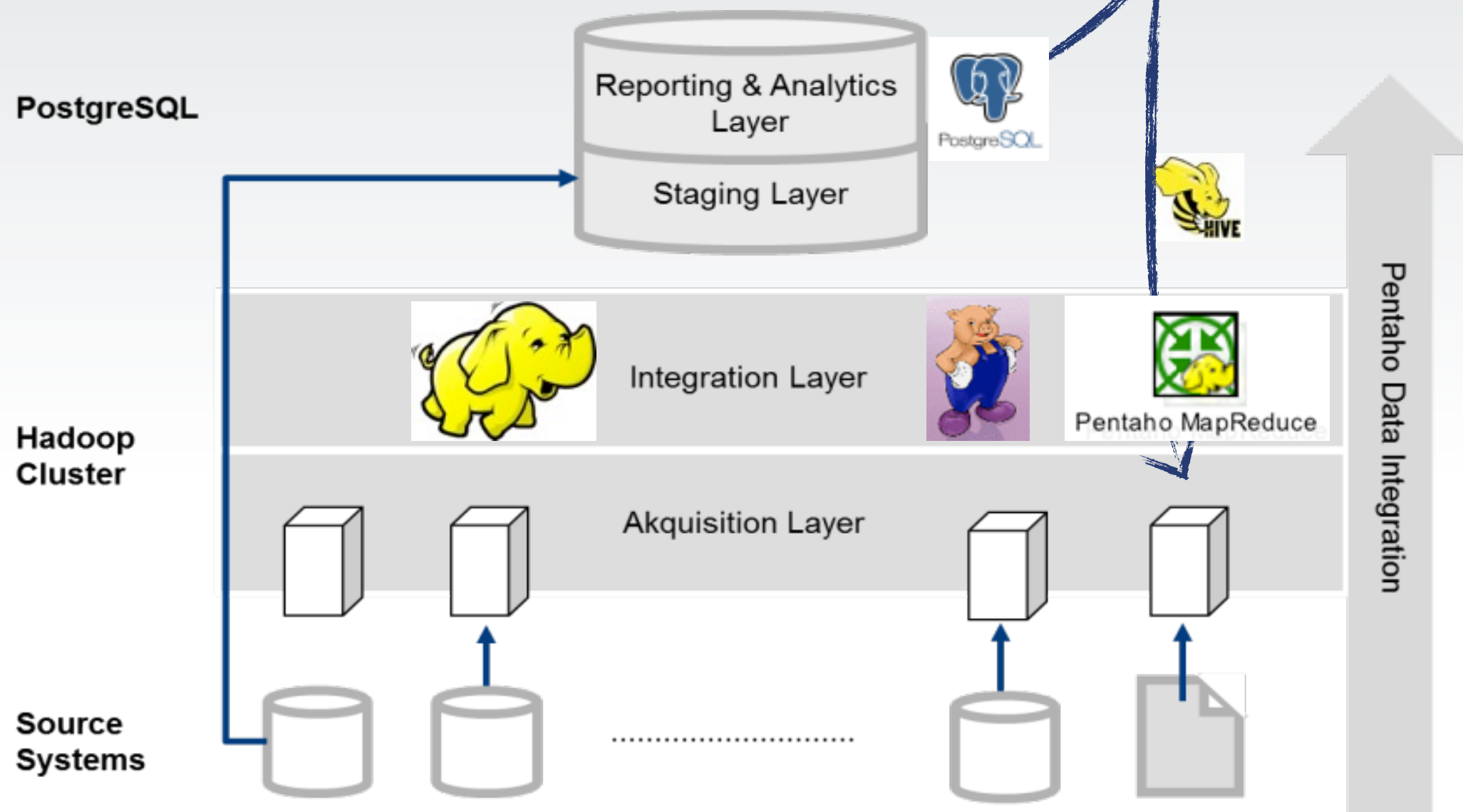
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster


inovex

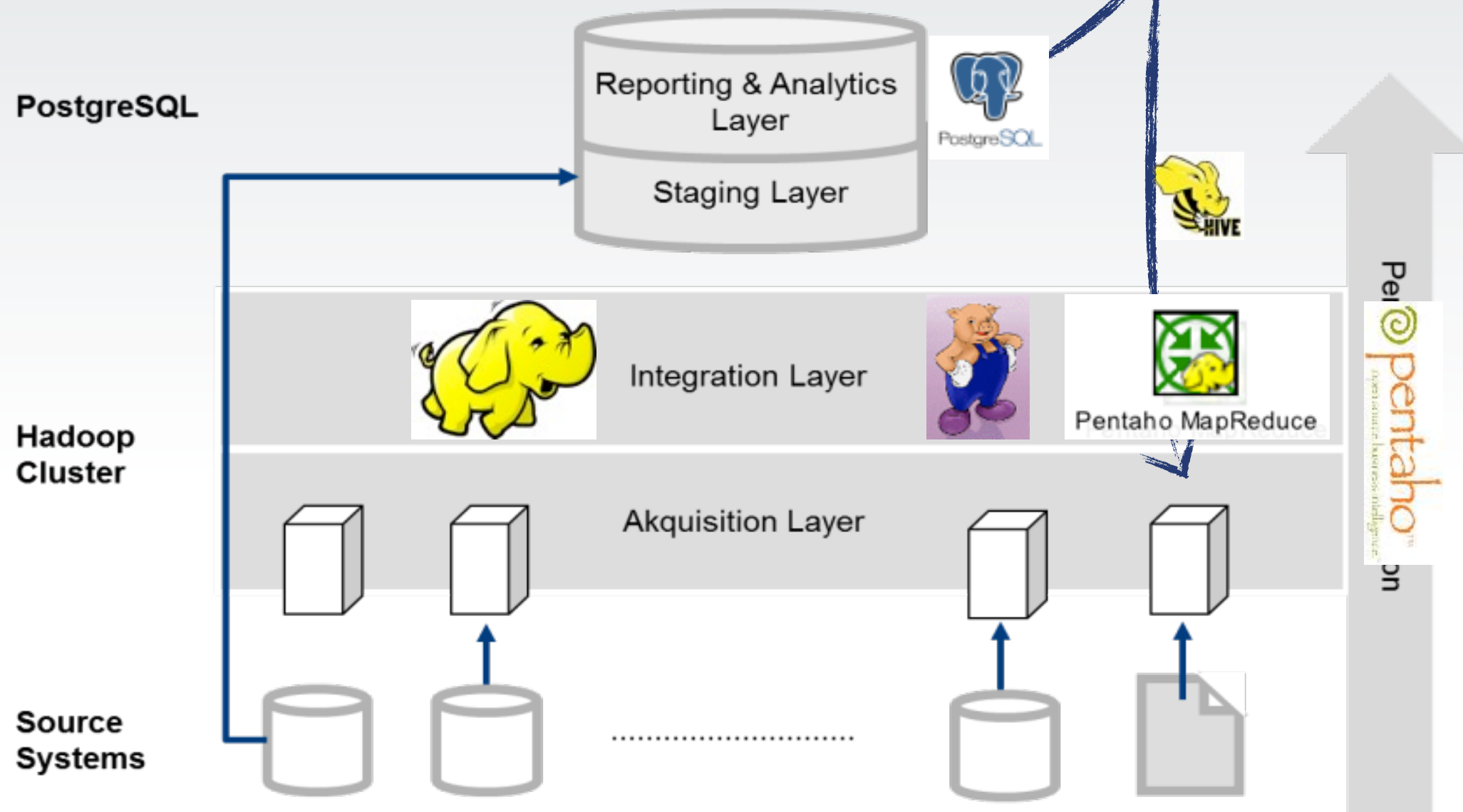
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster



Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster



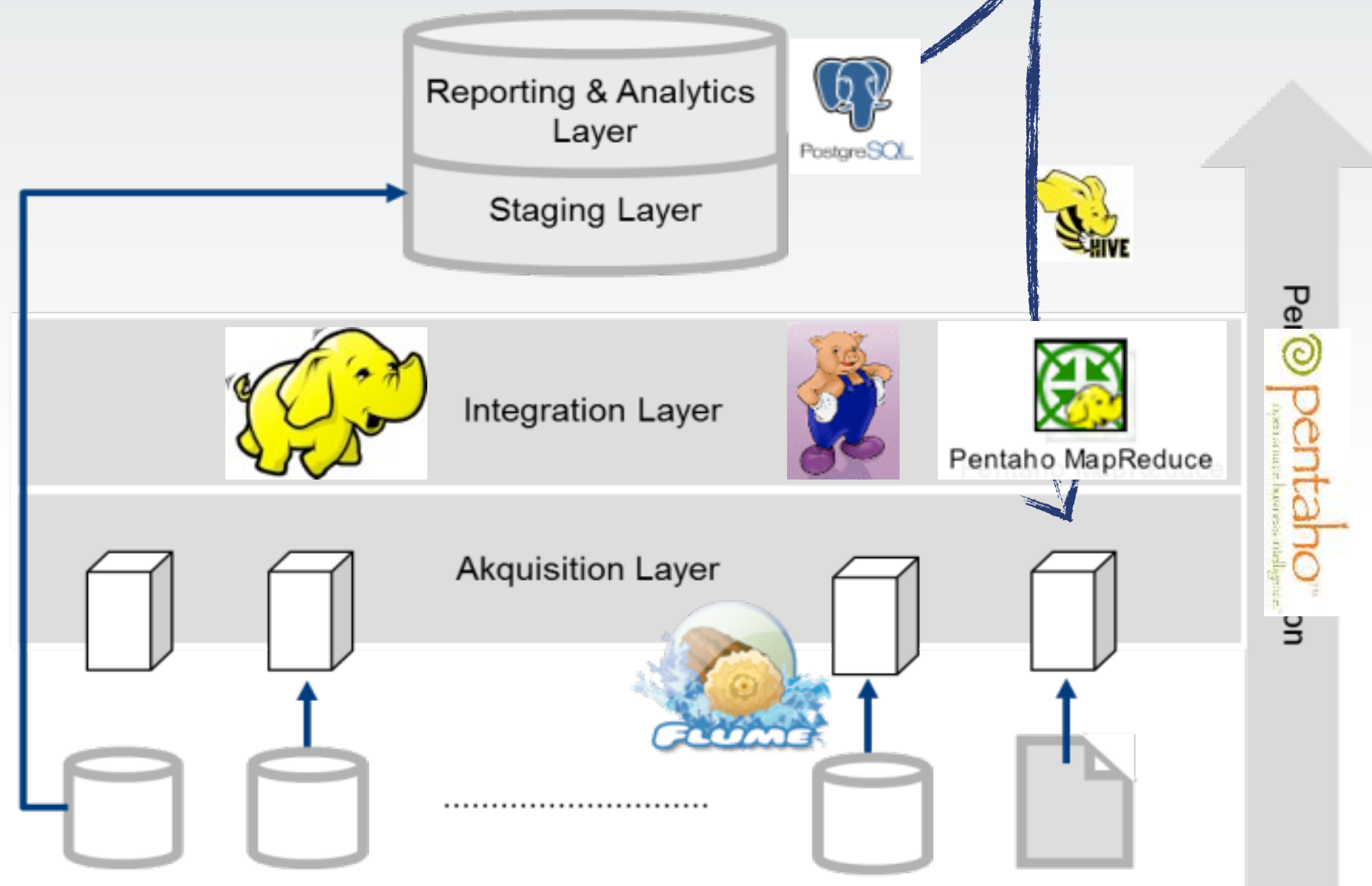
Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster

PostgreSQL

Hadoop Cluster

Source Systems



inovex

Werbung

1. U.a. **Hadoop-Entwickler-Training**

2. **1-3 Tage**

3. **Inhouse** oder **offen**

4. Offenes Trainings 2014:

18.-20. März (Köln)

24.-26. Juni (München)

18.-20. November (Karlsruhe)

www.inovex.de/trainings/offene-trainings/hadoop-training/

Fragen?