



SQL on Hadoop für praktikables BI auf Big Data

**Hans-Peter Zorn, Dr. Dominik Benz und
Sébastien Jelsch, Inovex GmbH**

War nicht BigData das gleiche
NoSQL?

Wie viele SQL Lösungen für Hadoop
gibt es mittlerweile?

Kylin
Oracle Hadoop Connector
Splice
InfiniDB
Hive on Spark
Pivotal HAWQ
Lingual
Shark
Hive-on-Tez
Tajo
Hive
Presto
BigSQL
Phoenix
Apache Drill
Impala
Trafodion
Hadapt
Spark SQL
CitrusDB
Aster SQL/Mapreduce
IBM BigSQL

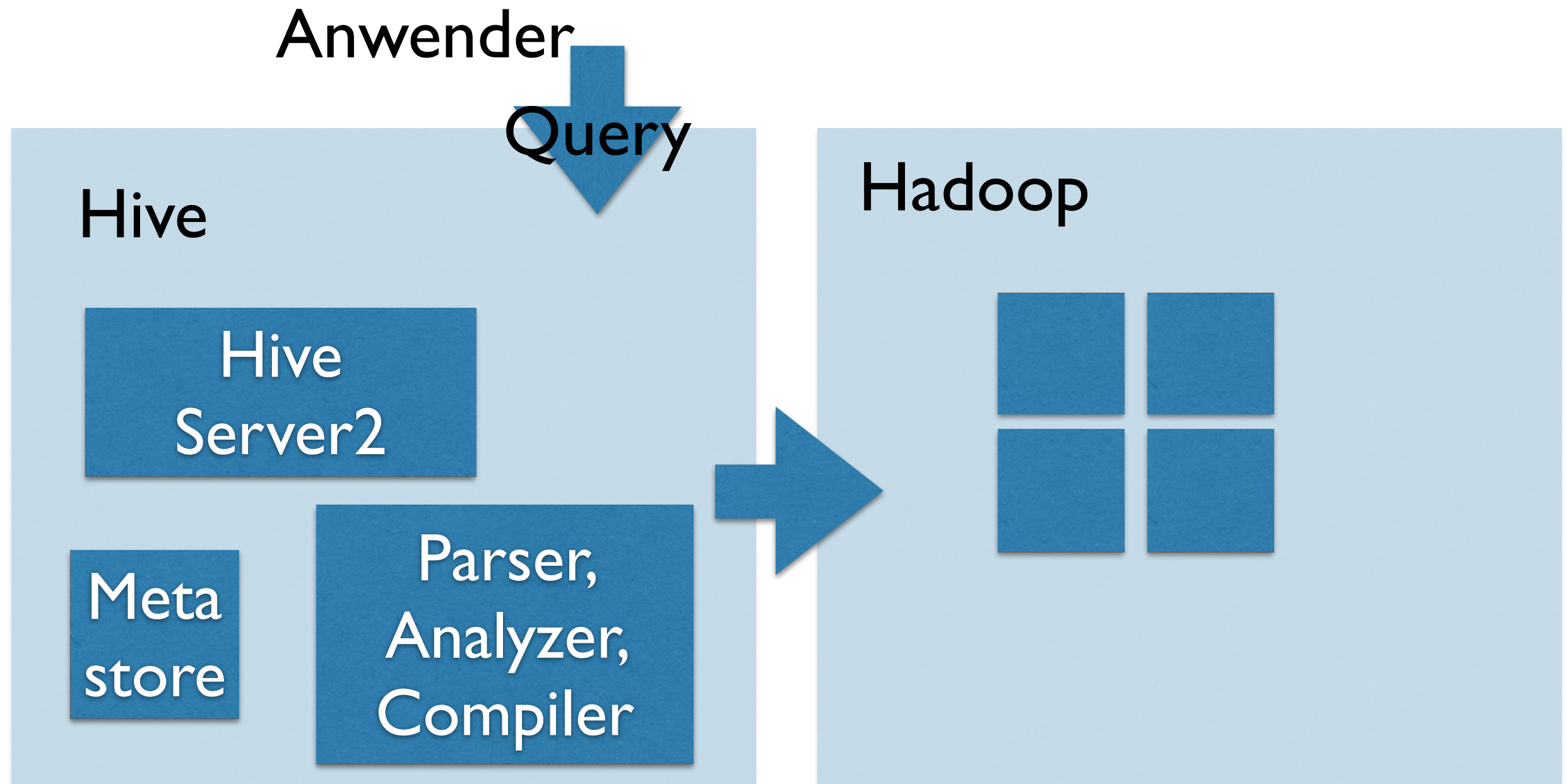
- Türöffner zur BigData-Welt
 - Kenntnisse weit verbreitet
 - Software:Treiber vorhanden



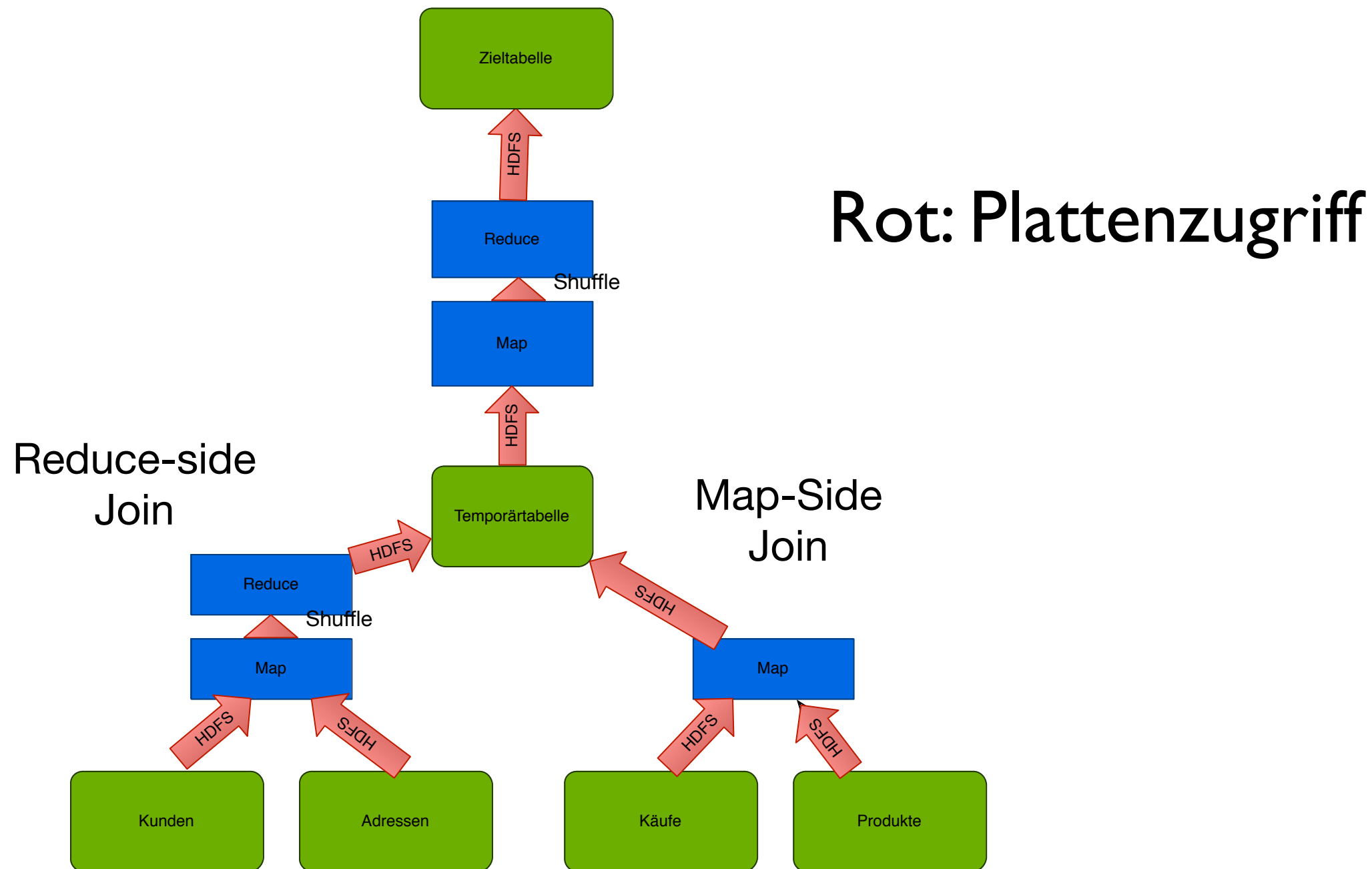
[flickr.com/salynaz](https://www.flickr.com/photos/salynaz/)

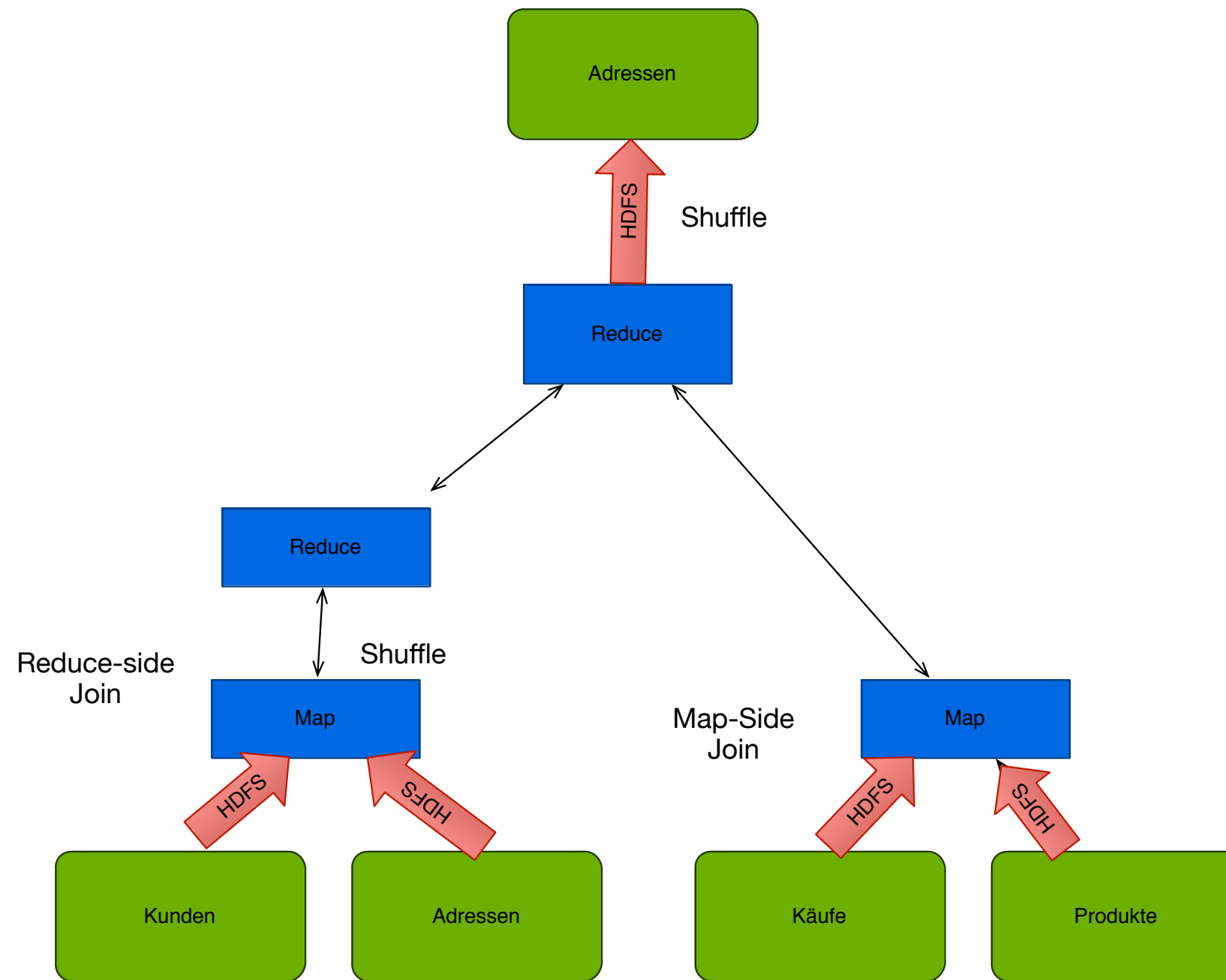
Oracle Hadoop Connector Splice InfiniDB
Kylin Hive on Spark Pivotal HAWQ
Shark Lingual
Hive-on-Tez Tajo **Hive** Presto
BigSQL Impala
Phoenix Apache Drill
Trafodion Hadapt Spark SQL
CitrusDB IBM BigSQL
Aster SQL/Mapreduce

- ursprünglich von Facebook
- Compiliert HiveQL, ein SQL-Dialekt, zu MapReduce-Jobs
- Schema wird separat zu den Daten abgelegt: Metastore

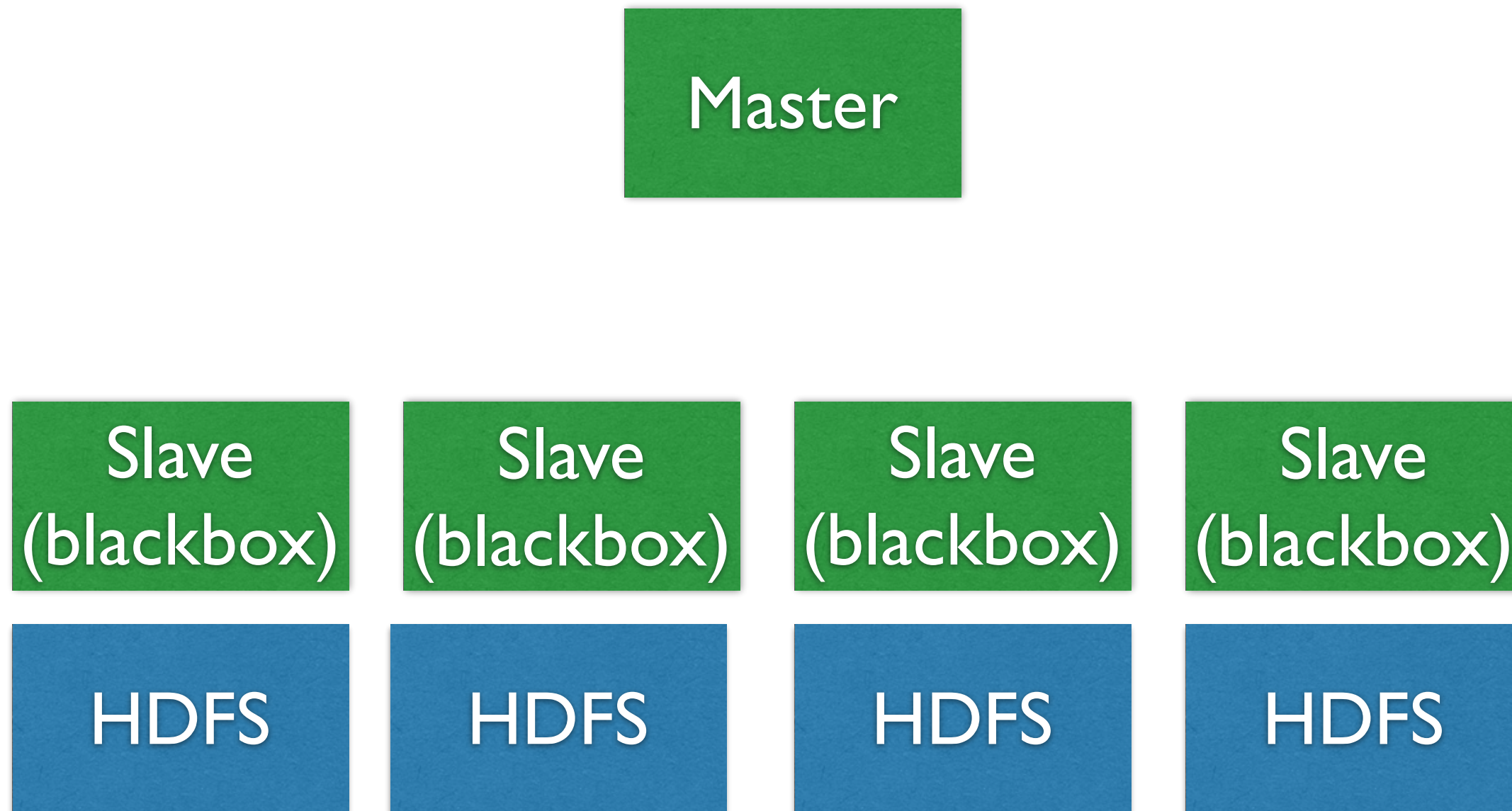


- ETL: gut geeignet
- Ad-hoc: Zu hohe Latenz
- Analytics: Latenz, Sprachumfang ungenügend



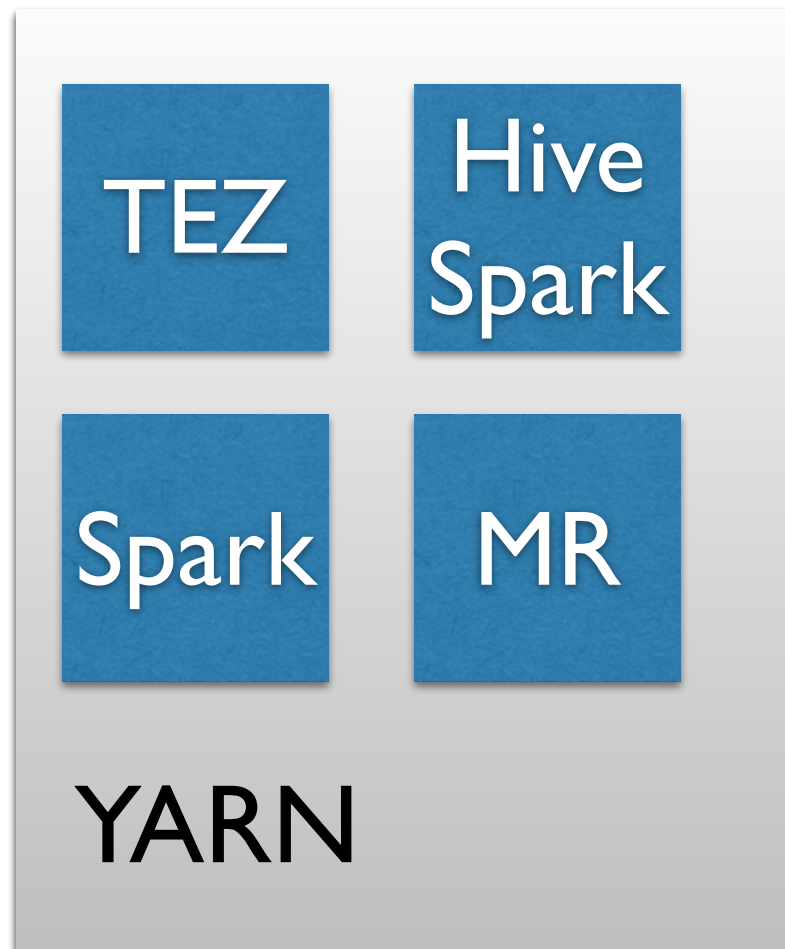


Oracle Hadoop Connector Splice InfiniDB
Kylin Hive on Spark Pivotal HAWQ
Shark Lingual
Hive-on-Tez Tajo Hive Presto
BigSQL Phoenix Apache Drill Impala
Trafodion Hadapt CitusDB Spark SQL
Aster SQL/Mapreduce IBM BigQuery



Oracle Hadoop Connector Splice InfiniDB
Kylin Hive on Spark Pivotal HAWQ
Shark Lingual
Hive-on-Tez Tajo **Hive** Presto
BigSQL Phoenix Apache Drill Impala
Trafodion Hadapt Spark SQL
CitusDB
Aster SQL/Mapreduce IBM BigSQL

- Hortonworks/Microsoft
- Hive weiterentwickeln
- Geschwindigkeit: Tez, ORC
- SQL-Features, Analytische Queries (OVER)
- Security (GRANT)



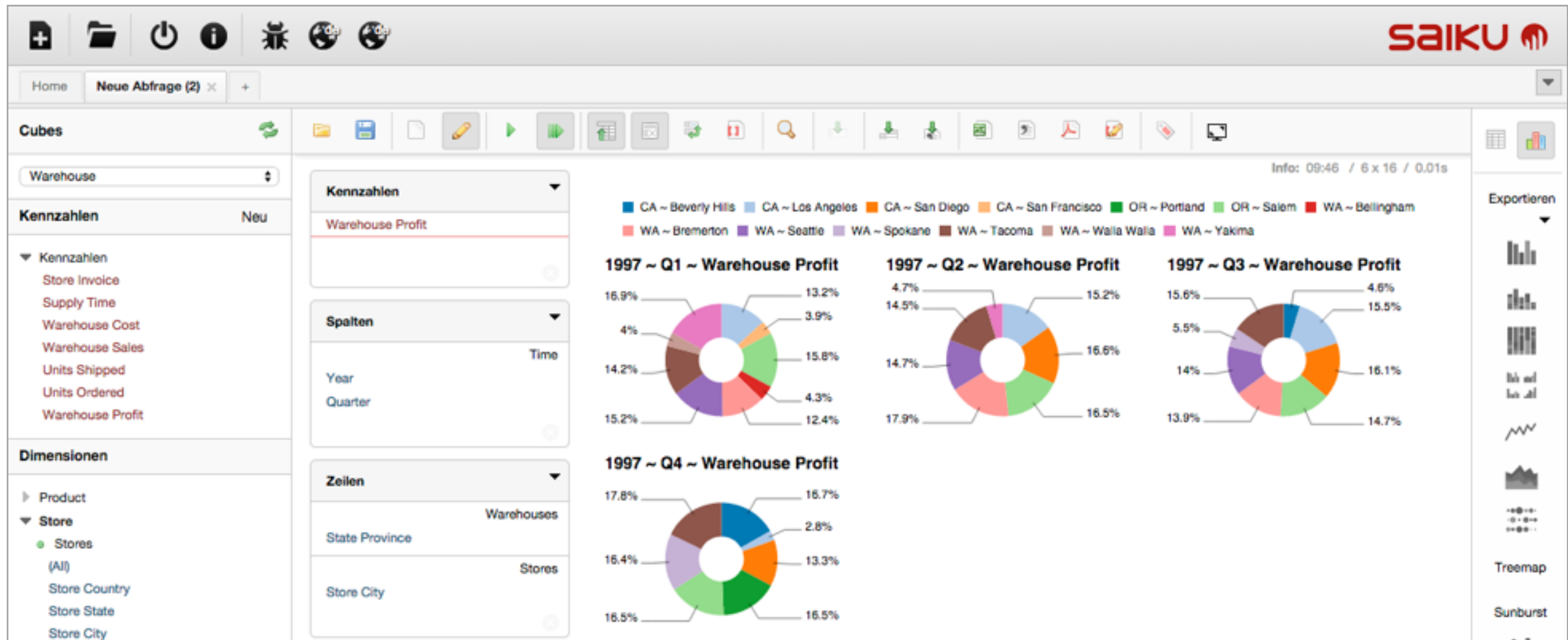
- YARN: Tez, Spark und Hive nebeneinander
- Tez: Neuer, spezialisierter
- Spark: generischer, viel Monumentum

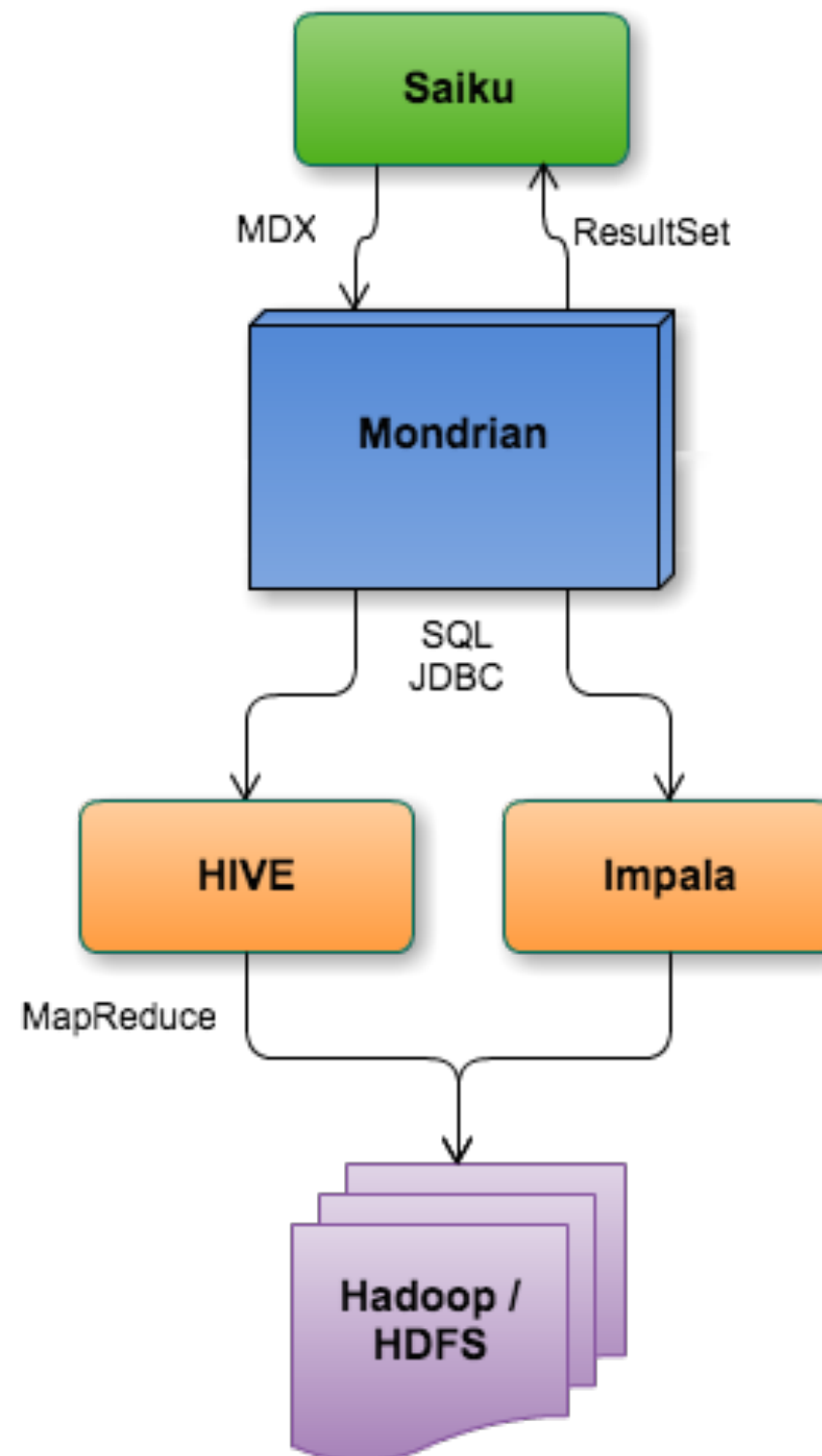
- MPP-basiert
- Queries zu nativem Code
- Speicherhungrig, empfohlen 128GB
- keine strukturierten Datentypen
- Zwischenergebnisse müssen in RAM passen (bis Impala 2.0)

- MPP-Engine
- Discovery-Server + Worker-Nodes
- Struktur-Datentypen -> JSON
- Hive, Cassandra, MySQL
- Anwendung bei Facebook: Fact-table in Hive, Dimensions in MySQL
- einfaches Deployment

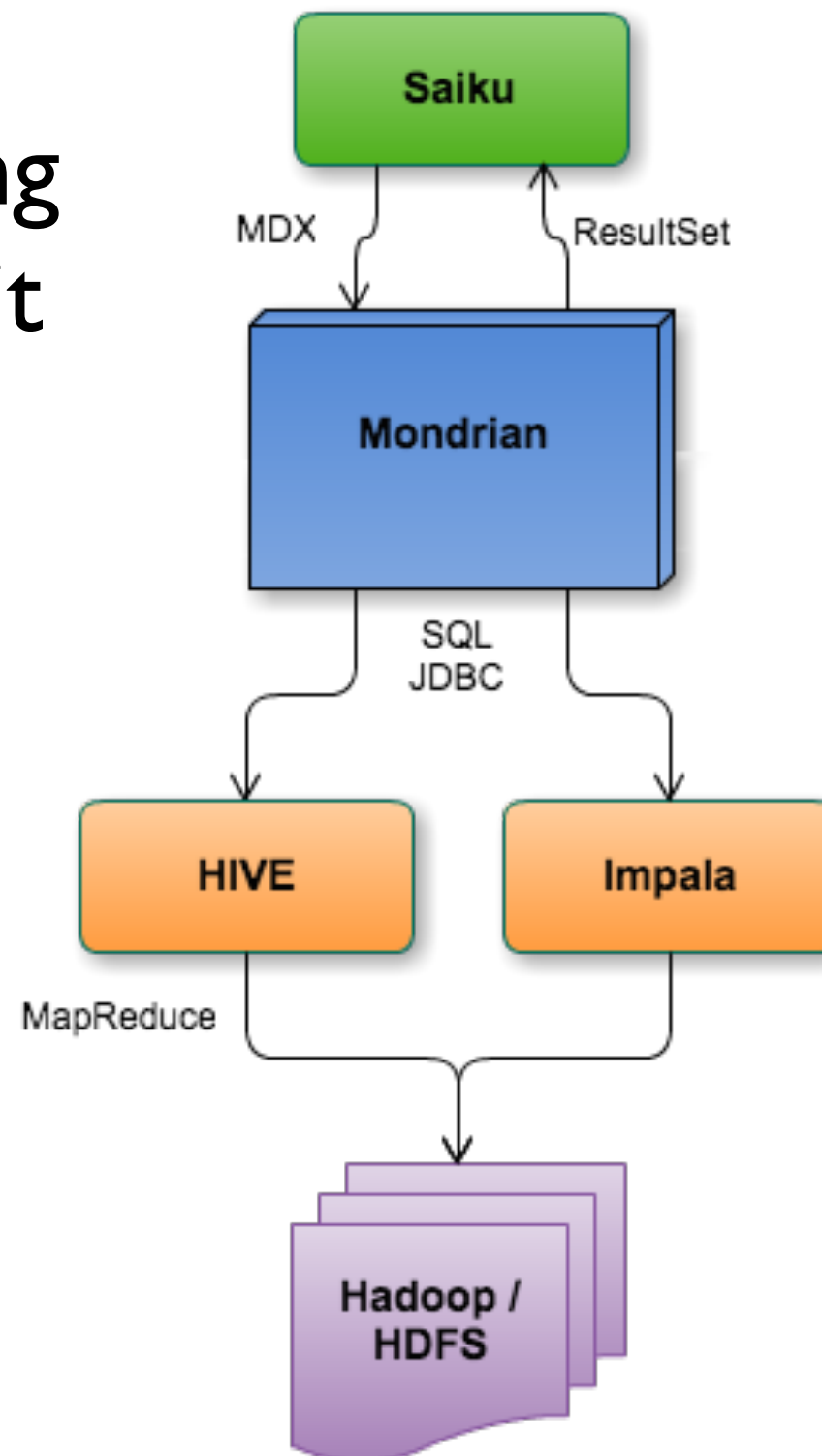
- MapR's
- Schema-on-Read
- Connectoren für Hive, HBase, JSON, CSV.
- Joins über mehrere Quellen hinweg
- Compiliert Queries zu Java-Byte-Code
- Version 0.5.0

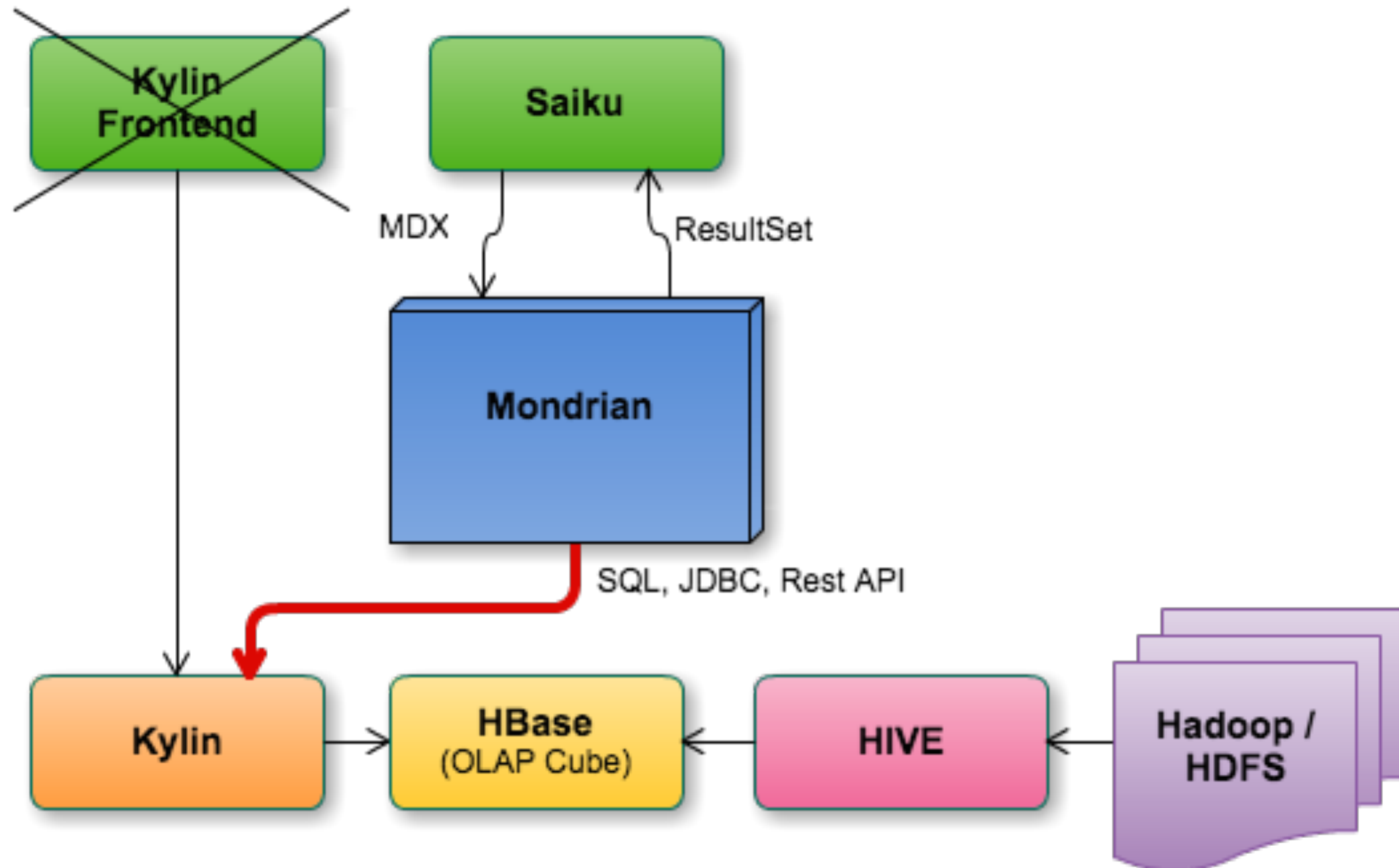
- Sehr neu, entwickelt von eBay
- (M)OLAP Engine
- Stern-Schema über Web-GUI definiert
- Aggregate werden per MapReduce vorgerechnet und in HBase gespeichert

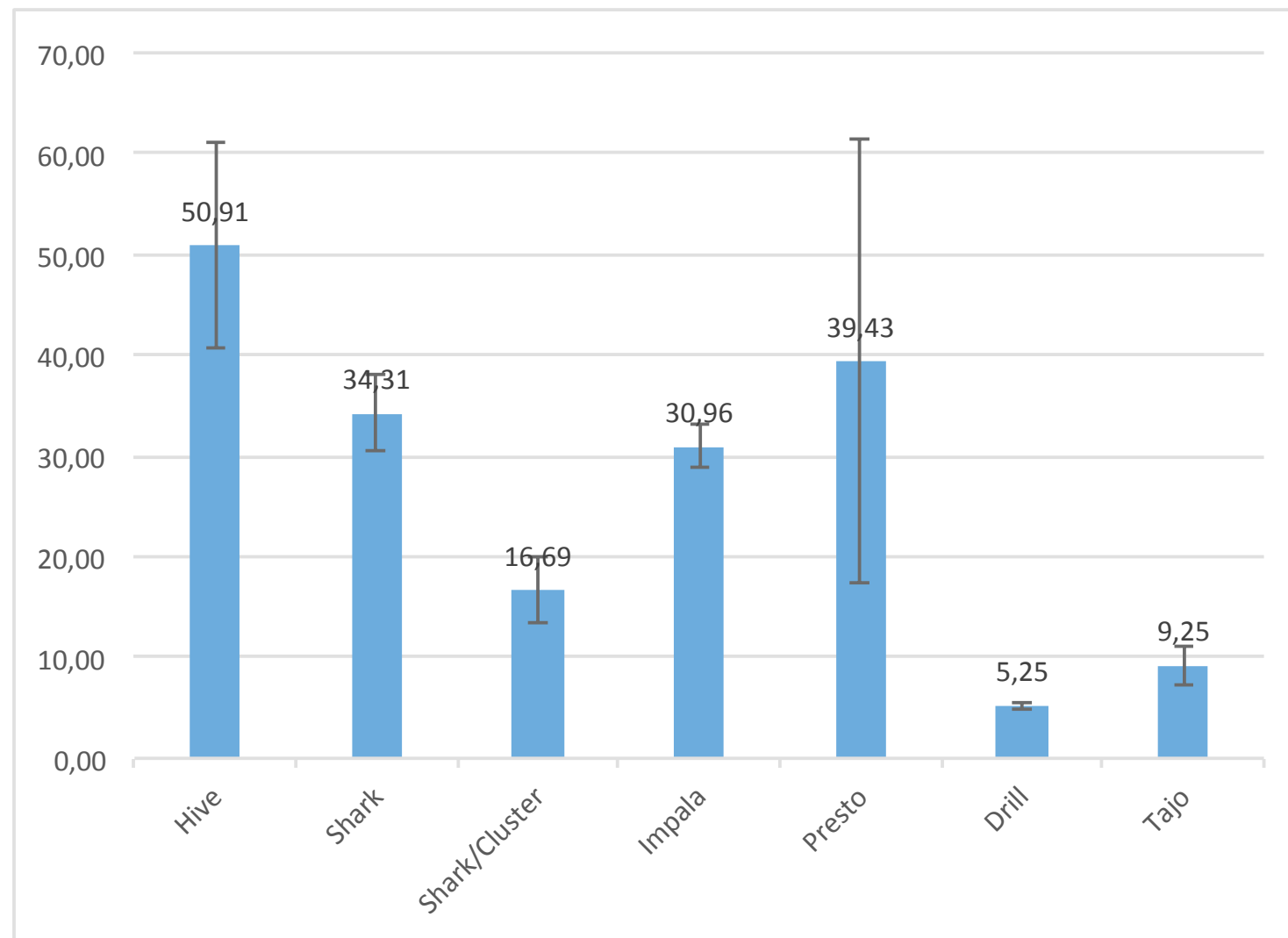


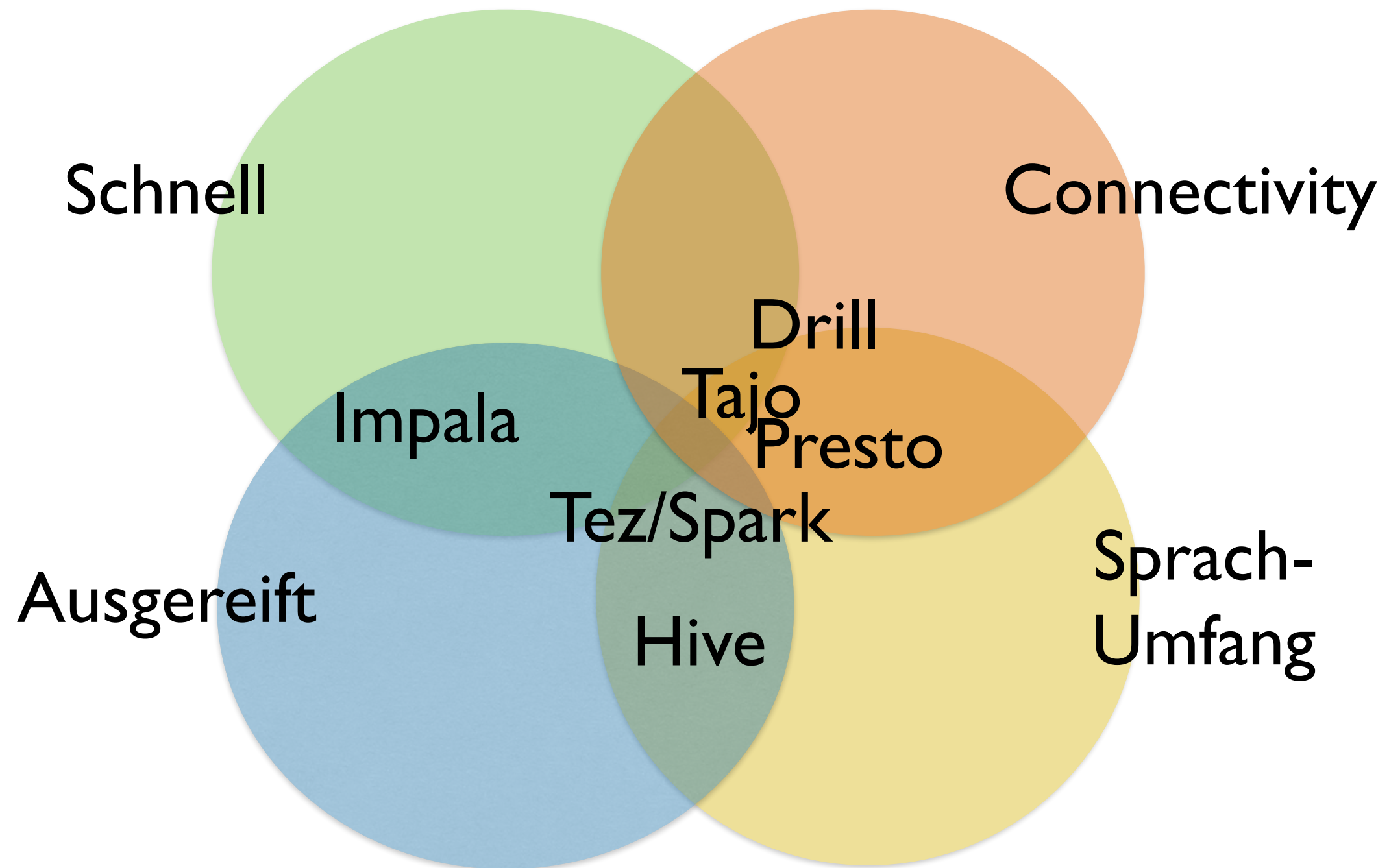


Praktische Erfahrung für dieses Setup mit Business Objects









Resumé

- Extrem schnell wachsendes, sich änderndes Umfeld.
- Keine „One-Size-Fits-All“ Lösung bisher
- Anhand des individuellen Anwendungsfalles zu evaluieren.
- Anbindung an BI-Tools wird besser

Vielen Dank für Eure Aufmerksamkeit!

Kontakt

Hans-Peter Zorn
Data Management & Analytics

inovex GmbH
Office Karlsruhe
Ludwig-Erhard-Allee 6
76131 Karlsruhe

0173 31 81 093
hzorn@inovex.de

