Big Data zum Anfassen

Spark, Hive, Kafka & Co.

Wolfgang Pleus – **PLEUS** Consulting

www.pleus.net

IT Solution Architect , Entwickler, Agile Coach

> 20 Jahre Projekterfahrung

> 10 Jahre agile Projekte

DAX - Startup

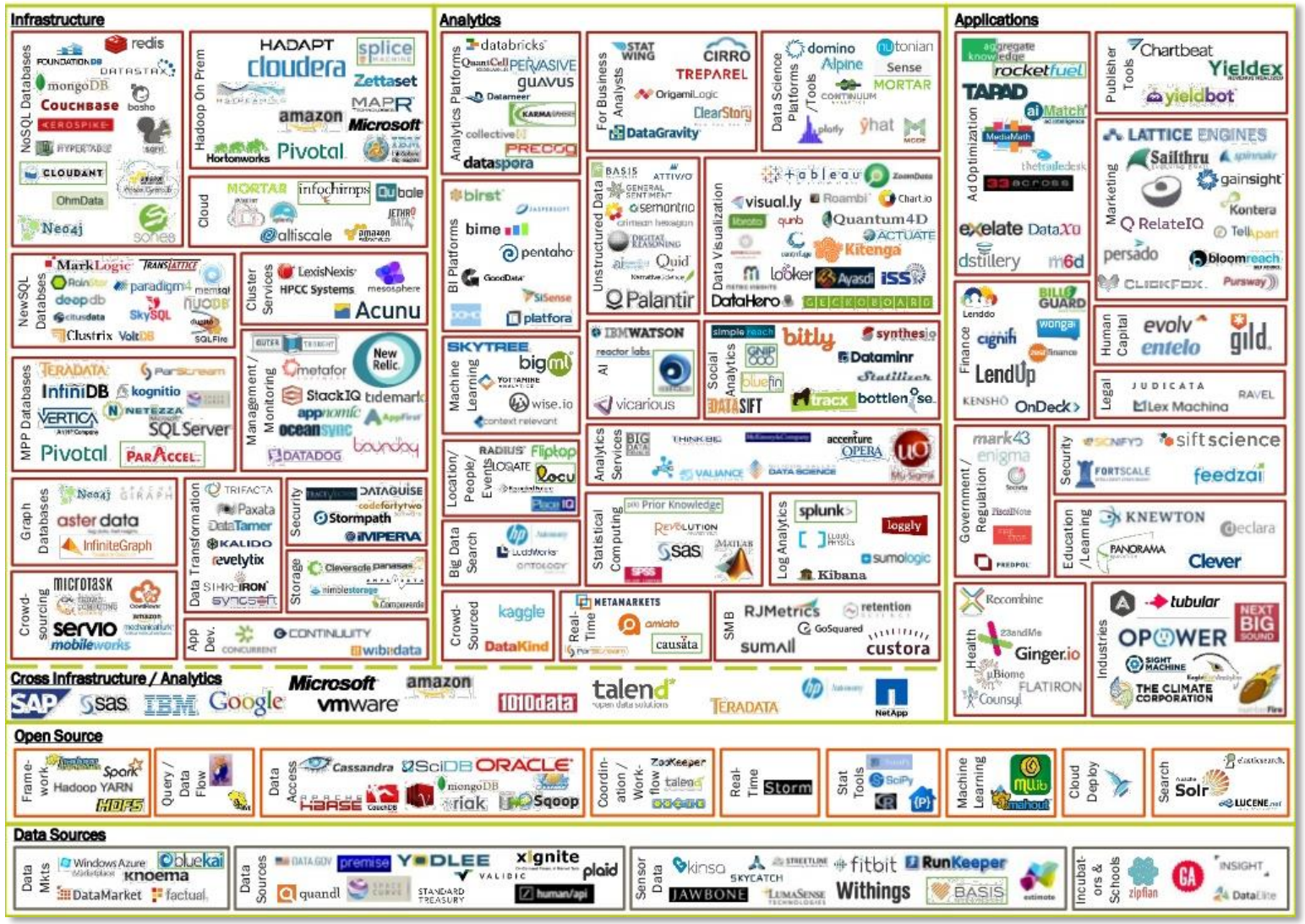
Wolfgang Pleus – **PLEUS** Consulting

www.pleus.net

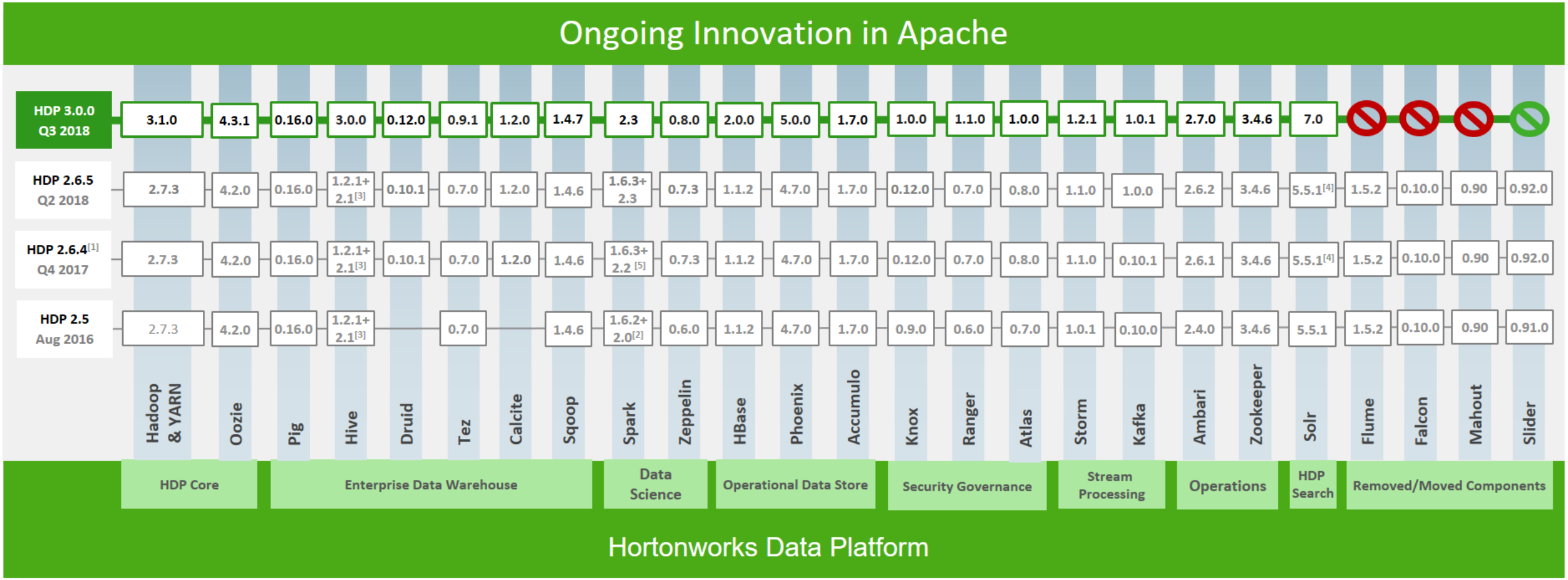
Agenda

- Big Data Stack
- Small Data
- Lambda Architektur
- Lauffähiger Prototyp
- Grüne Wiese on Premise

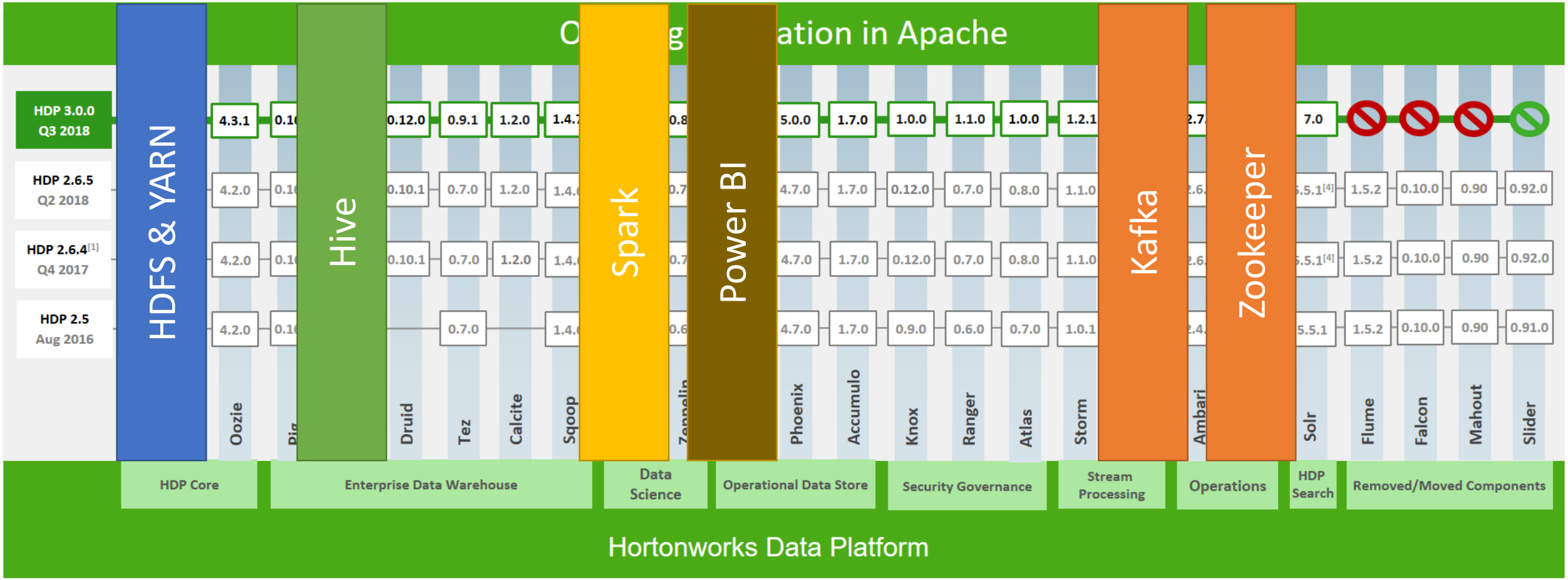
Big Data Ökosystem



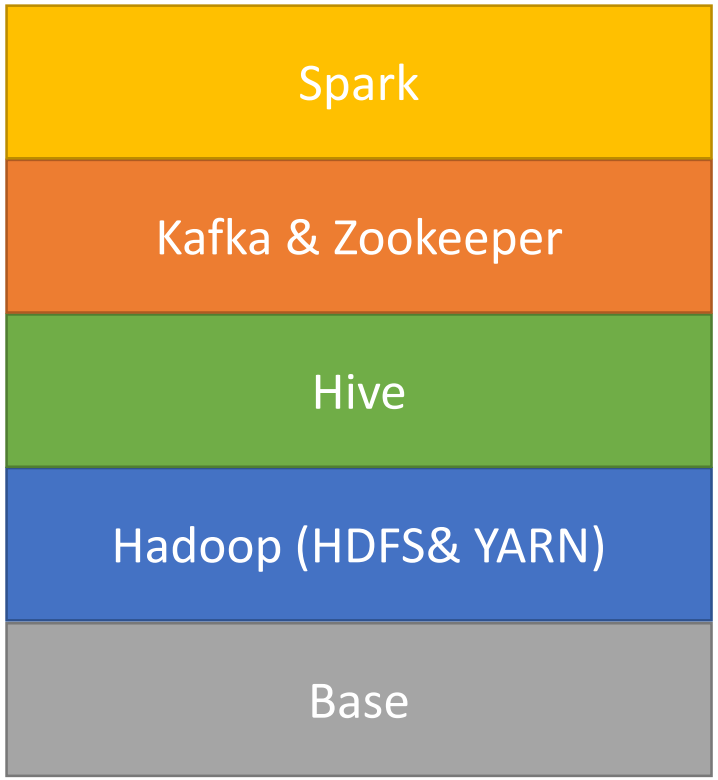
Distribution



Minimales Setting

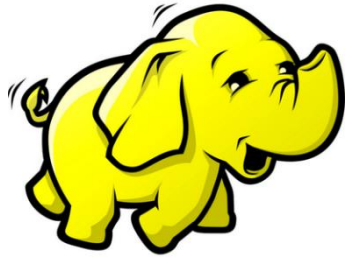


Docker Images

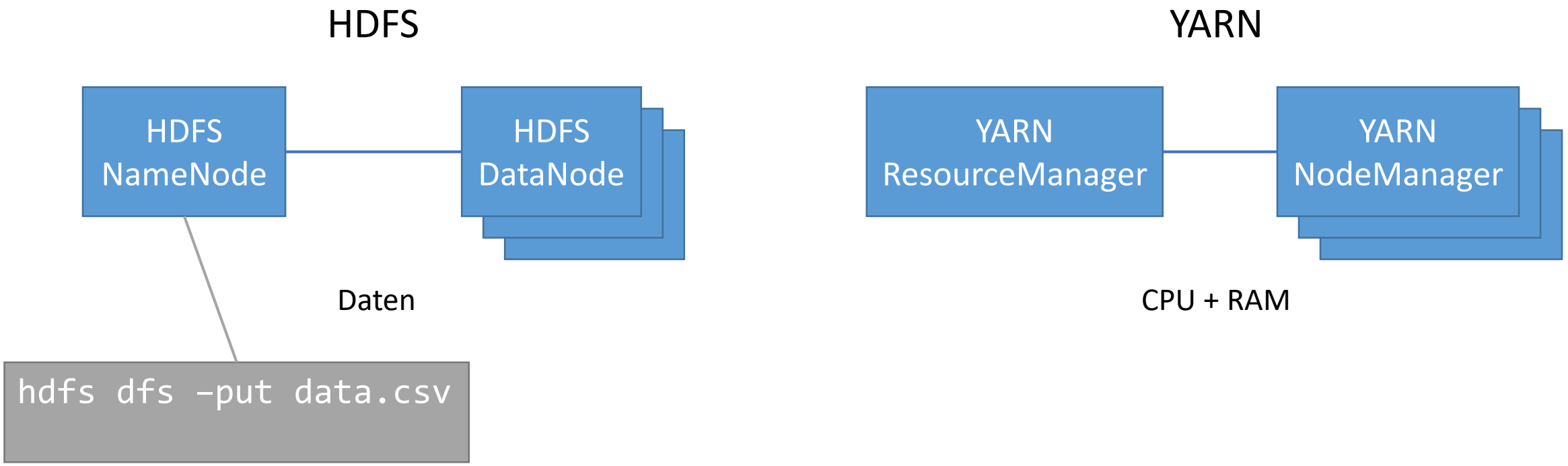


Demo: Basis

Hadoop: HDFS & YARN



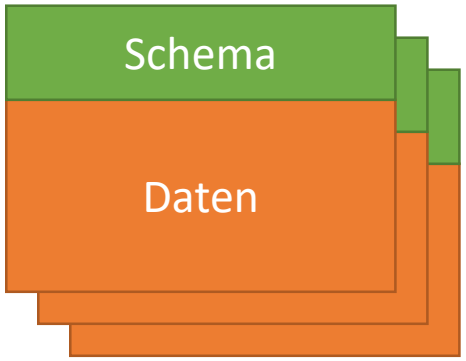
- Skalierbarer Speicher, CPU und Memory



Demo: Hadoop

Avro

- Schema für Datasets
- Serialisierung
- Schemaevolution



jobs.avro

```
{"namespace": "net.pleus.bddemo",
 "name": "jobs",
 "version":"1.0",
 "doc" : "Jobs",
 "type": "record",
 "fields": [
     {"name": "created", "type": "string","default":"unknown", "doc":"Record creati
     {"name": "published",  "type": ["string", "null"],"default":"unknown","doc":"P
     {"name": "source", "type": ["string", "null"],"default":"unknown","doc":"Sourc
     {"name": "title", "type": ["string", "null"],"default":"unknown","doc":"Name o
     {"name": "type", "type": ["string", "null"],"default":"unknown","doc":"Offer t
     {"name": "link", "type": ["string", "null"],"default":"unknown","doc":"Link to
     {"name": "region", "type": ["string", "null"],"default":"unknown","doc":"Work
     {"name": "geolocation", "type": ["string", "null"],"default":"unknown","doc":"
     {"name": "longitude", "type": ["string", "null"],"default":"unknown","doc":"Wo
     {"name": "latitude", "type": ["string", "null"],"default":"unknown","doc":"lat
 ]
}
```

jobs.avsc

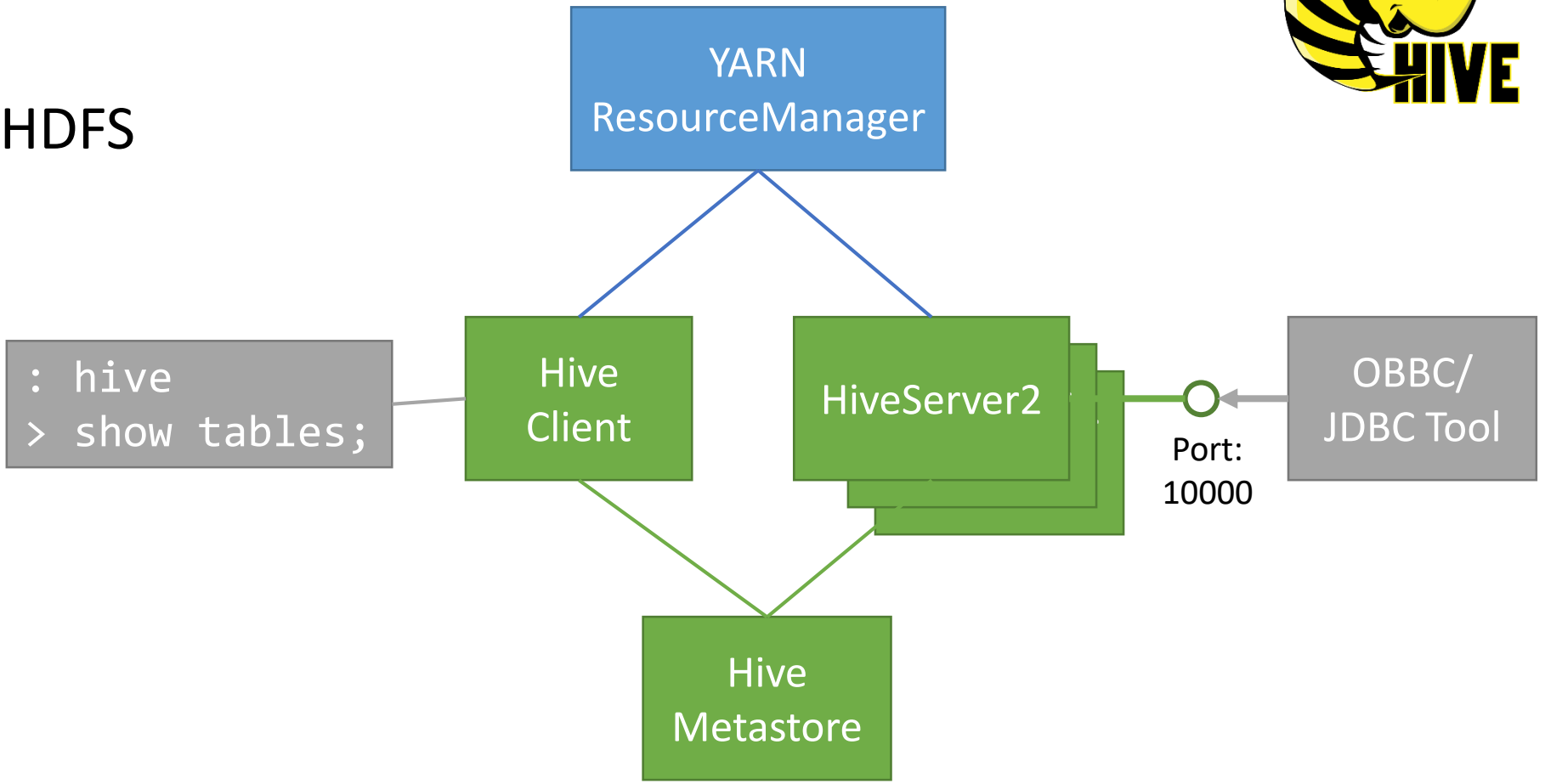
Sqoop

- SQL + Hadoop
- Import und Export relationaler Daten
- Paralleles Laden mit Map Reduce

```
sqoop import
--connect jdbc:oracle:thin:@server:1521:demo
--username wpl
--password=***
--create-hive-table
--table ORA.JOBS
-m 1
--hive-import
--hive-overwrite
--hive-table jobs
```

Hive

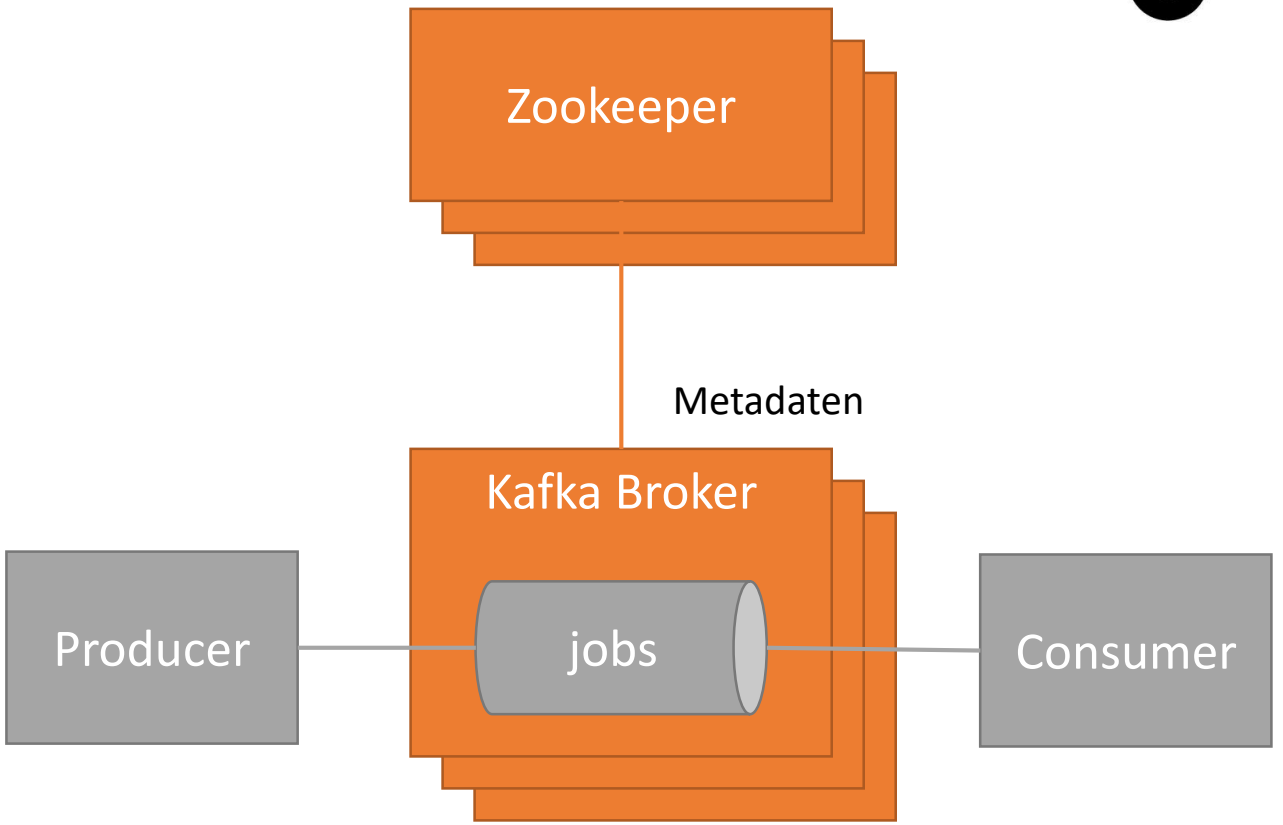
- SQL Layer für HDFS



Demo: Hive

Kafka

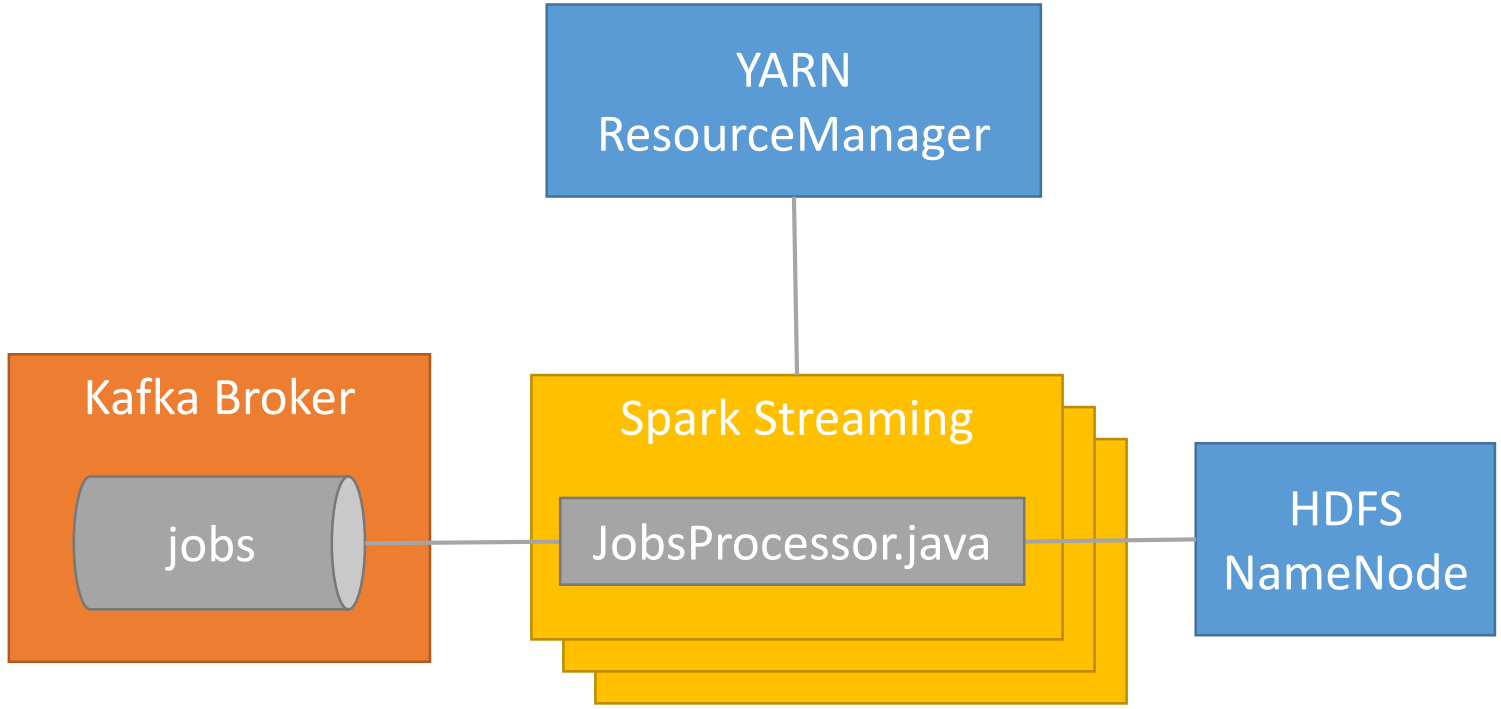
- Message Broker
- Streamverarbeitung
- Skalierbar



Demo: Kafka

Spark

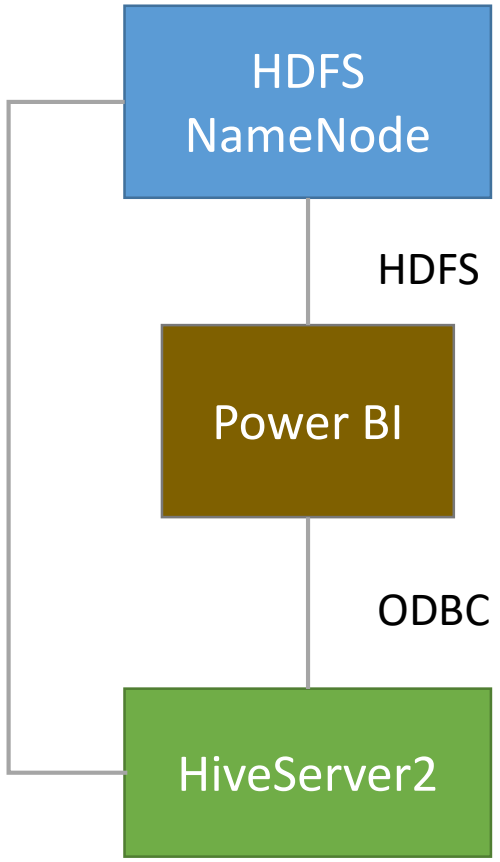
- Streaming
- SQL
- Transformationen
- Clusterfähig



Demo: Spark

Visualisierung

- Z.B. Power BI

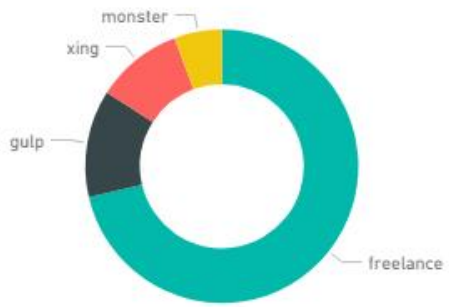


Demo: Visualisierung

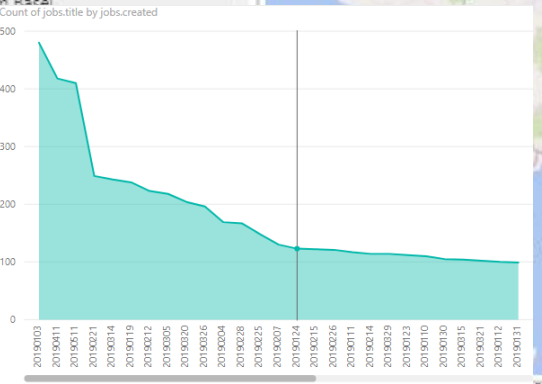
Count of jobs.source by jobs.source

5005

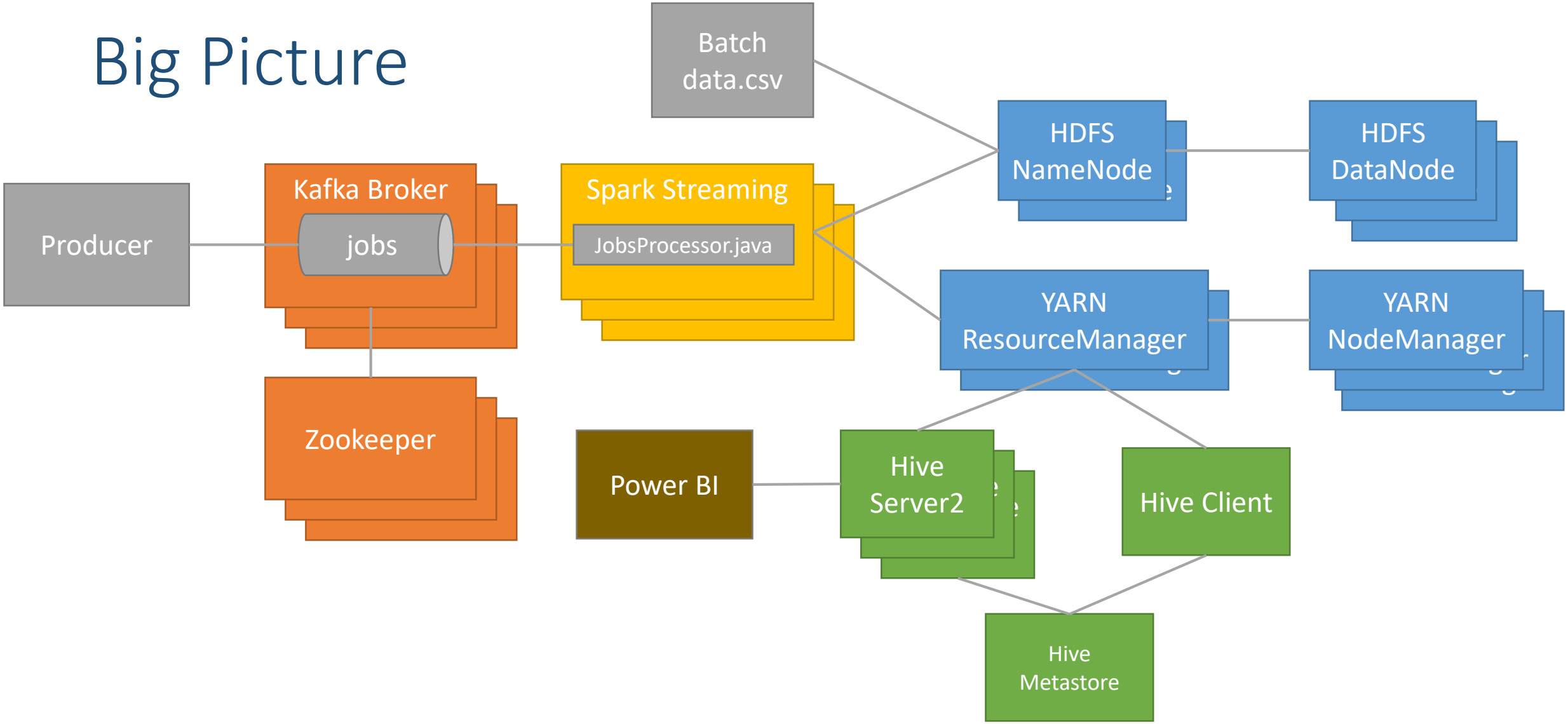
Count of jobs.title



jobs.title	jobs.created	jobs.region
ILE RPG - Entwickler (m/w/x)	20190103	1445 Luxemburg
Agile Full-Stack Developer (m/w)	20190103	A-Wien
Angular2+-Webentwickler (m/w) 393254/6	20190103	A-Wien
FREELANCER4VIENNA - Salesforce Vlocity Developer (f/m)	20190103	A-Wien
MS Dynamics Entwickler	20190103	A-Wien
Entwickler - webbasierte Anwendungen fÃ¼r eine Robo Advisor-Plattform (m/w)	20190103	Berlin
Java Entwicklung (w/m)	20190103	Bern
Project Manager in the Scientific and Manufacturing field	20190103	CH-Aarau
Agile .NET Developer 100% (m/w) 394208/19	20190103	CH-Basel
Python Developer / Business Analyst - Data Modelling/SQL Server	20190103	CH-Basel
Java/Angular Entwickler (m/w), Pos. 1756	20190103	CH-Bern
Network Engineer / Migration Expert (w/m)	20190103	CH-Bern
Professional Software Quality Specialist and Integration Developer (m/f) 393789/11	20190103	CH-Großraum Basel
Business Analyst (m/w)	20190103	CH-Großraum
Senior Testautomatisierer (m/w), Pos. 1757	20190103	CH-Großraum
Scrum Master / Scrum Coach (5866)	20190103	CH-Großraum
DevOps Engineer (100%) - C#/Powershell/HyperV/SCVMM /Lucerne	20190103	CH-Großraum
Applikations Architekten/Entwickler – WebService (60%, 80% & 100%) (5961)	20190103	CH-Großraum
Cobol Entwickler - zOS / DB2 (5946)	20190103	CH-Großraum
IT Business Analyst – Agile (5935)	20190103	CH-Großraum
IT Business Analyst / RE - Kollektivleben (5948)	20190103	CH-Großraum
Java Backend Developer (5699)	20190103	CH-Großraum
Java Entwickler (5712)	20190103	CH-Großraum
Scrum Master (m/w)	20190103	CH-Großraum



Big Picture


Batch
data.csv
HDFS
NameNode
HDFS
DataNode
Producer
Kafka Broker
jobs
Spark Streaming
JobsProcessor.java
YARN
ResourceManager
YARN
NodeManager
Zookeeper
Power BI
Hive
Server2
Hive Client
Hive
Metastore

Fazit

- Der Einstieg ist einfach
- Wirtschaftlich und technisch skalierbar
- Ausprobieren lohnt sich

Vielen Dank

www.pleus.net

wolfgang.pleus@pleus.net

Download Folien und Code www.pleus.net/blog

Creative Software Workbench www.cswob.de

Lambda Architektur