

Koexistenz?

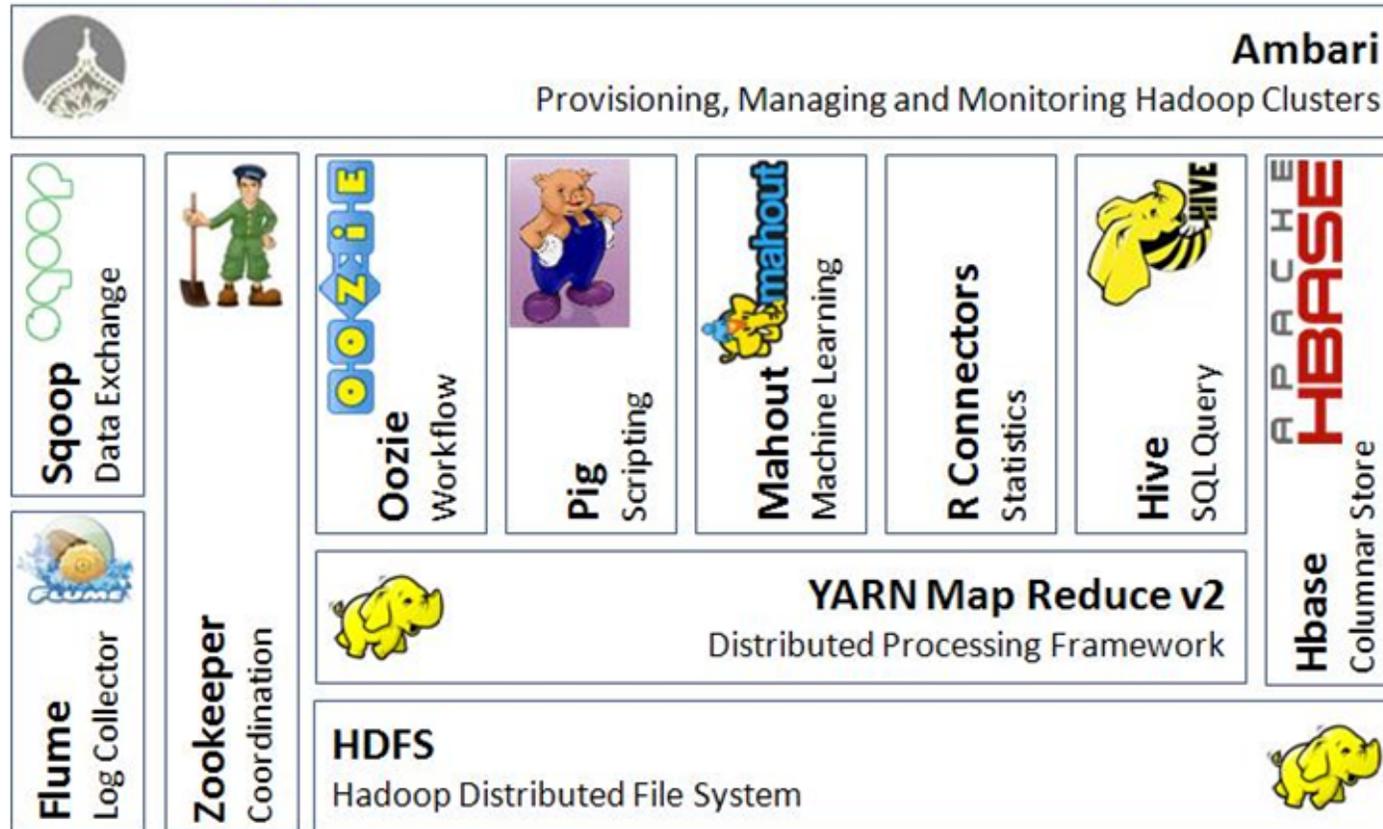
Oracle und Hadoop
gemeinsam nutzen.

Matthias Jung
DOAG Regio 10.09.2019



- Das Hadoop-Ökosystem
- Oracle Big Data Connectors (Auswahl)
- Demonstration: SQL Connector für HDFS

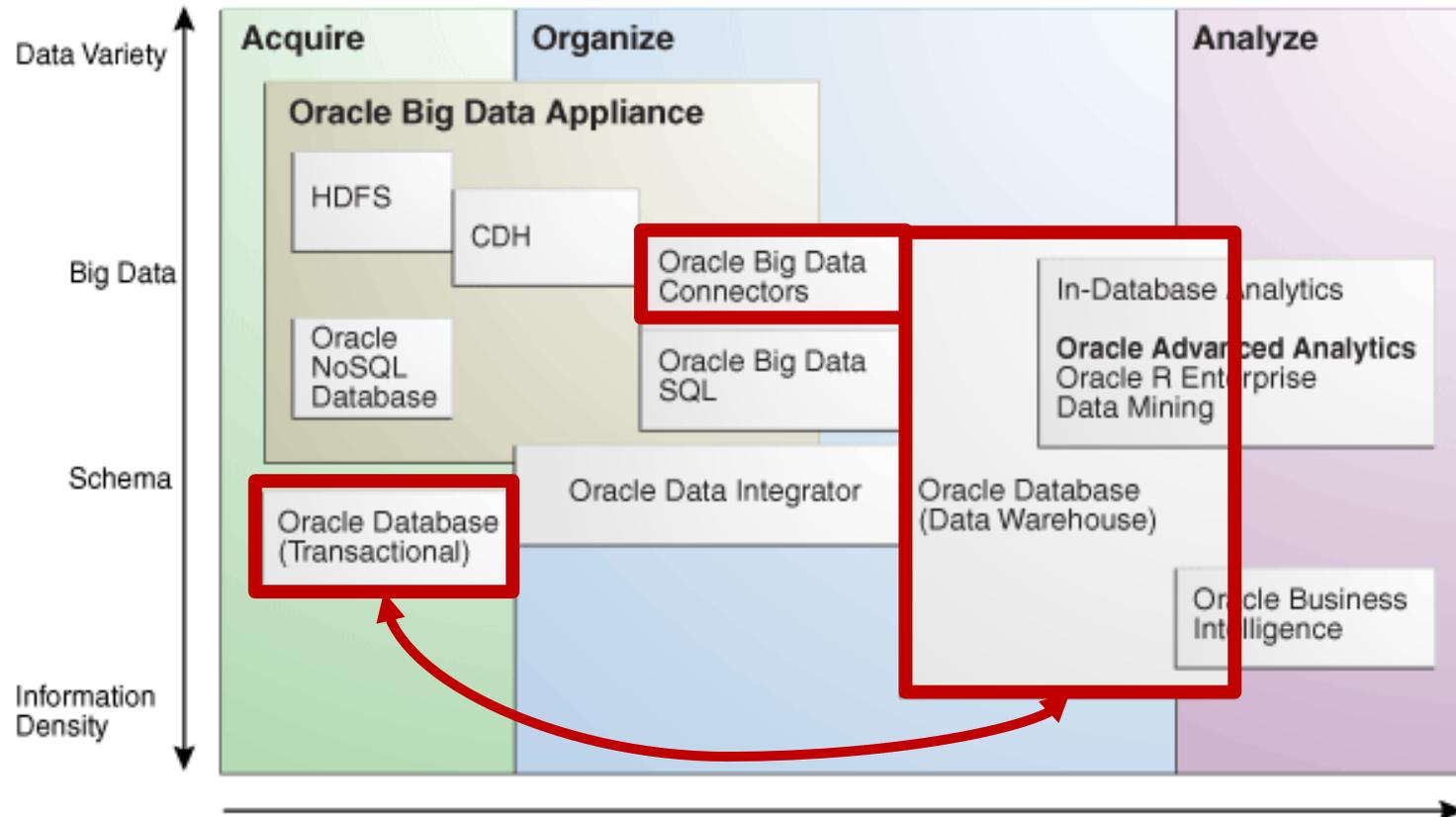
Hadoop Ecosystem



<https://www.facebook.com/hadoopers>

Note: This is not an exhaustive list

Das Oracle Big Data Ökosystem



https://docs.oracle.com/cd/E55905_01/doc.40/e55814/concepts.htm#BIGUG119

Worum geht es?

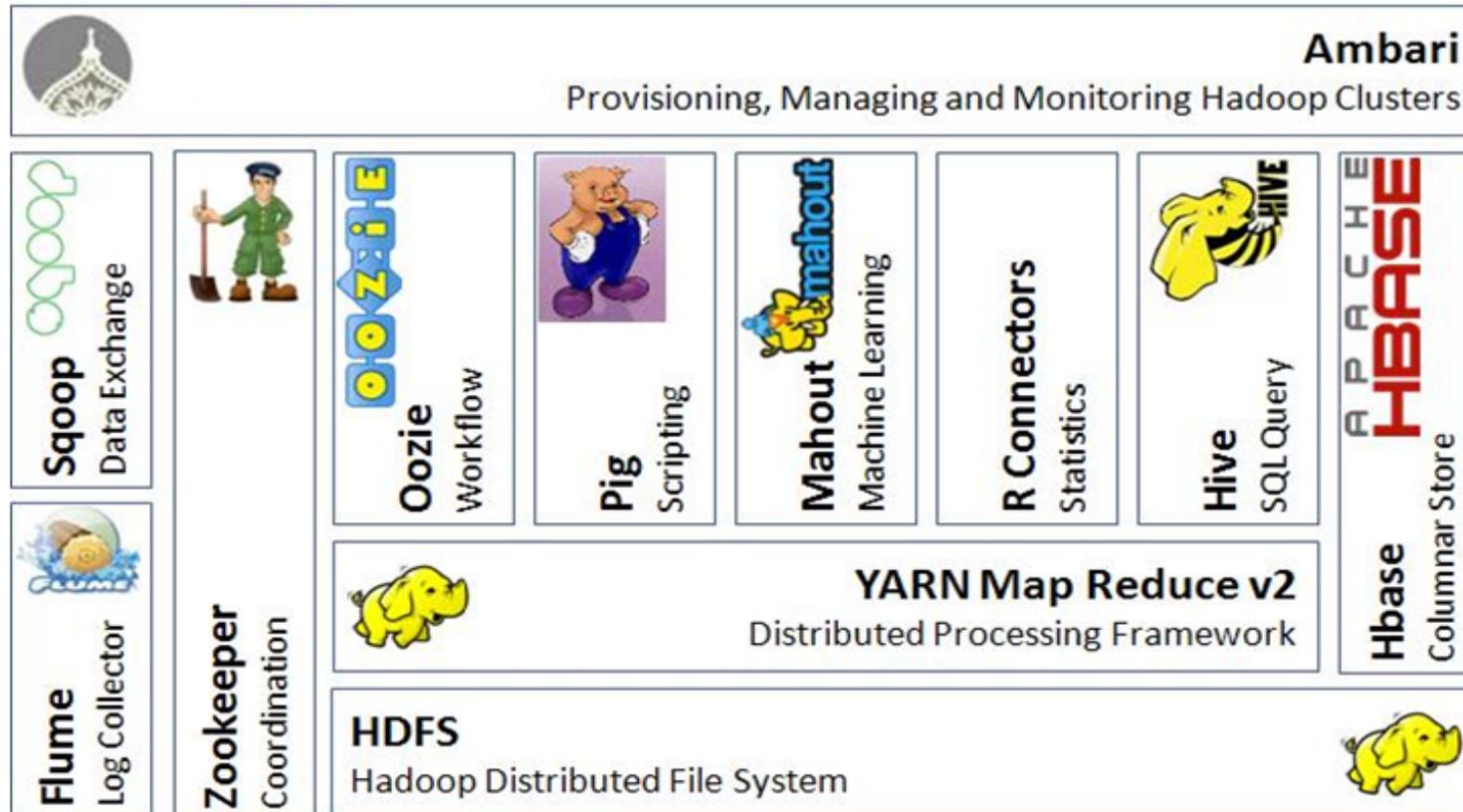
- Interaktion zwischen einer „normalen“ Oracle DB und Hadoop (HDFS)
- Worum geht es nicht:
 - Oracle Big Data Appliance / Oracle Big Data SQL
 - Oracle GoldenGate
 - Oracle Database Gateway for ODBC
 - Oracle Data Integrator (ODI)
 - Oracle Advanced Analytics
 - Sonstige ETL Produkte

Agenda

- Das Hadoop-Ökosystem
- Oracle Big Data Connectors (Auswahl)
- Demonstration: SQL Connector für HDFS

- Oracle Data
- Oracle Log
- Oracle R A
- Oracle XQ
- Oracle Dat
- **Oracle SQ**

Hadoop Ecosystem



Note: This is not an exhaustive list

<https://www.facebook.com/hadoopers>

The screenshot shows a news article on the website 'bigdata-insider.de'. The article title is 'Cloudera und Hortonworks vereinigen ihre Datenplattformen'. The main text discusses the merger of Cloudera and Hortonworks into a single entity, Cloudera, and the introduction of a new Enterprise Data Cloud. The article mentions that this happened at the DataWorks conference in Barcelona in January 2019. The article is dated 08.04.19 and is written by Stefan Girschner and Nico Litzel. There are social media sharing buttons for Twitter, Facebook, LinkedIn, PDF, and Print. A URL is provided at the bottom of the article: <https://www.bigdata-insider.de/cloudera-und-hortonworks-vereinigen-ihre-datenplattformen-a-817696/>

Cloudera und Hortonworks vereinigen ihre Datenplattformen

LIVE | 12.09.2019 | 10:00 Uhr
WEBINAR ANMELDUNG ▶

CLUDERA DATA PLATFORM

Unified control plane
Public, private & hybrid clouds
Shared data experience
Powered by open source
Analytics from the Edge to AI

Altus DataPlane
Identity | Orchestration | Management | Operations

Data Engineering
Data Warehouse
Operational Database
Machine Learning

sdX
Catalog | Schema | Migration | Security | Governance

Nachbericht DataWorks Summit Barcelona
Cloudera und Hortonworks vereinigen ihre Datenplattformen
08.04.19 | Autor / Redakteur: Stefan Girschner / Nico Litzel

Die Cloudera Data Platform (CDP) enthält mehrere Analyse-Frameworks wie DataFlow & Streaming, Data Engineering, Data Warehouse, Operational Database und Machine Learning. (Bild: Cloudera)

Auf der diesjährigen DataWorks-Konferenz in Barcelona traten Cloudera und Hortonworks erstmals gemeinsam auf. Im Januar 2019 hatten beide auf Data Science und Big Data spezialisierten Anbieter ihren Zusammenschluss vollzogen. Als ein Ergebnis wird demnächst die neue Enterprise Data Cloud eingeführt, die vollständig auf Open Source basiert.

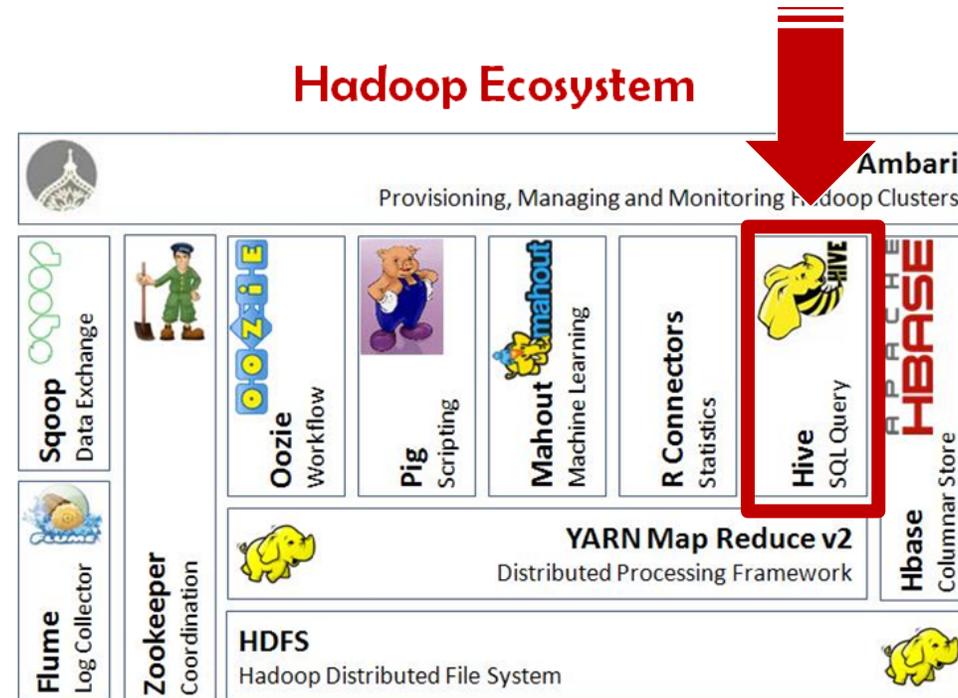
Als Cloudera vor zehn Jahren gegründet wurde, gingen die Gründer davon aus, dass die Unternehmen ihre Daten in der Cloud speichern, verwalten und analysieren möchten. So entstand der Name Cloudera, ein Wortspiel aus Cloud und Ära. Aber die Mehrheit der Unternehmen waren noch nicht bereit für die Cloud, sie wollten die Datenplattform vor Ort nutzen. So kam es, dass Cloudera zunächst On-premises-Lösungen für das Speichern, Verwalten und Analysieren von Daten angeboten hat. Heute gibt es kaum noch Vorbehalte, eine Datenplattform aus der Cloud nutzen.

share me
share me
tweet me
share me
PDF
Weiterempfehlen
Drucken

<https://www.bigdata-insider.de/cloudera-und-hortonworks-vereinigen-ihre-datenplattformen-a-817696/>

Oracle Datasource for Apache Hadoop

- Hive QL- / Spark SQL-Zugriff auf Oracle-Tabellen
- Oracle dient als Quelle
- Daten in Hadoop können mit Oracle Daten „gejoined“ werden
- Auf Seiten der Oracle DB ist nicht viel zu tun:
 - Bereitstellen der Tabellenstruktur
 - User-Account



Note: This is not an exhaustive list

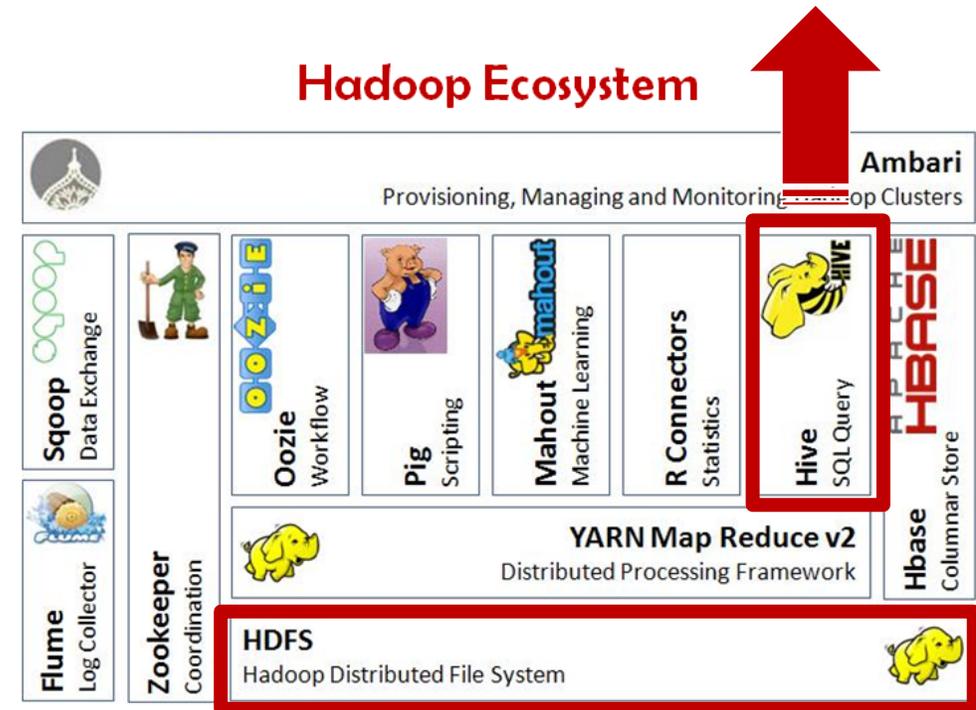
<https://www.facebook.com/hadoopers>

- Beispiel:

```
hive> CREATE EXTERNAL TABLE ora_tab
      STORED BY 'oracle.hcat.osh.OracleStorageHandler' TBLPROPERTIES (
        'mapreduce.jdbc.url' = 'jdbc:oracle:thin:@//oraclehost:1521/prod',
        'mapreduce.jdbc.username' = 'scott',
        'mapreduce.jdbc.password' = 'tiger',
        'mapreduce.jdbc.input.table.name' = 'employee',
        'oracle.hcat.osh.useMonitor'='true',
        'oracle.hcat.osh.fetchSize'='10000',
        'oracle.hcat.osh.useOracleParallelism'='true'
      );
```

Oracle Loader for Hadoop

- „ETL“-Tool, welches in der Lage ist verschiedene Hadoop-Quellen in eine Oracle DB zu laden
- Hadoop ist die Quelle
- Formate
 - Text Files (z.B. aus HDFS)
 - Hive-Tabellen
 - Parquet
 - JSON
- Zusatzfunktionen
 - Datentransformationen (Typen)
 - Kompression (in der DB)
- Online- / Offline-Modus



Note: This is not an exhaustive list

<https://www.facebook.com/hadoopers>

Online Database Mode

- Ablauf
 - OLH verbindet sich zur Laufzeit mit der Zieldatenbank
 - Metadaten der Zieltabelle werden gelesen
 - Daten werden aus Hadoop gelesen und direkt in die Zieltabelle geschrieben
- Vorteile
 - Einfach zu konfigurieren
 - Einfach zu betreiben
- Nachteile
 - Verbindung von Hadoop zur Datenbank notwendig
 - Höhere Last auf Oracle als im Offline Mode
 - Nicht so schnell wie Offline mit Data Pump-Dateien

Offline Database Mode

- Ablauf
 - Metadaten der Zieltabelle werden aus Datei gelesen
 - Daten werden aus Hadoop gelesen
 - Daten werden in HDFS-Datei geschrieben (Data Pump-Datei)
- Vorteile
 - Zur Laufzeit keine Verbindung zur Datenbank notwendig
 - Offline mit Data Pump-Dateien ist am schnellsten
 - Geringste Last auf der Oracle-Datenbank
- Nachteile
 - Aufwendiger zu konfigurieren
 - Datei Transfer zusätzlich notwendig

- **Beispiel:**

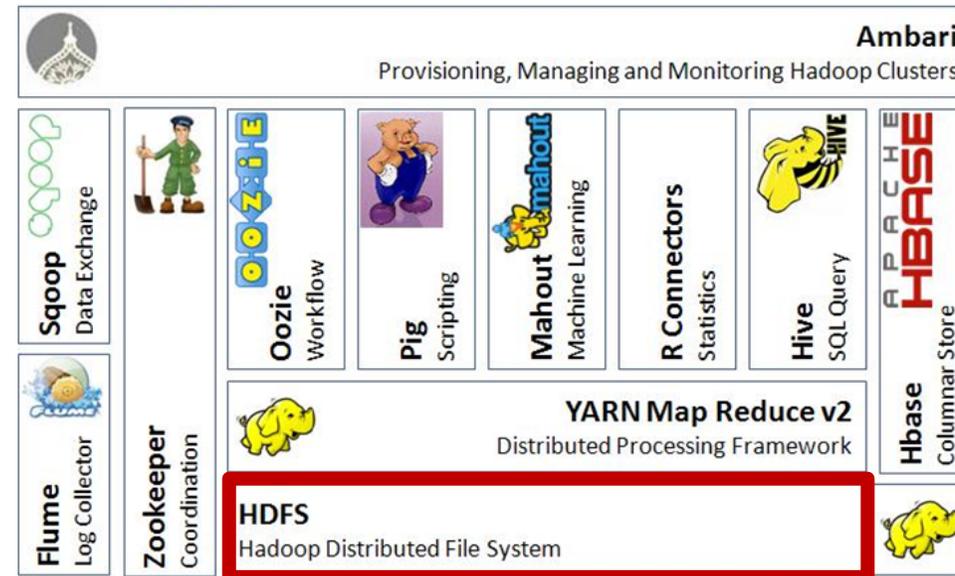
(<https://blogs.oracle.com/bigdataconnectors/how-to-load-oracle-tables-from-hadoop-tutorial-part-2-hello-world>)

```
bash> $HADOOP_HOME/bin/hadoop jar $OLH_HOME/jlib/oraloader.jar
oracle.hadoop.loader.OraLoader
-D oracle.hadoop.loader.jobName=OLHP_fivdti_dtext_jdbc_0_722
-D oracle.hadoop.loader.loaderMapFile=file:/tmp/loaderMap_fivdti.xml
-D mapred.reduce.tasks=0
-D mapred.input.dir=/user/olh_performance/fivdti/56000000_90
-D mapred.output.dir=/user/oracle/olh_test/results/fivdti/722
-conf /tmp/oracle_connection.xml
-conf /tmp/dtextInput.xml
-conf /tmp/jdbcOutput.xml      (Online: Direkt über JDBC in die DB)
```

Oracle SQL Connector for HDFS

- SQL Interface auf HDFS-Dateien
- HDFS dient als Quelle für die Abfrage (SELECTs)
- „Read-access“
- Zugriff wird über „external tables“ realisiert
- Kann zum Laden von Daten in die DB genutzt werden

Hadoop Ecosystem



Note: This is not an exhaustive list

<https://www.facebook.com/hadoopers>

Agenda

- Das Hadoop-Ökosystem
- Oracle Big Data Connectors (Auswahl)
- Demonstration: SQL Connector für HDFS

SQL Connector for HDFS

- Installation
- Hadoop Client auf der Datenbank installieren
[Cloudera YUM Repository installieren]
`yum install hadoop-client`
- HDFS-Konfiguration von HDFS Name Node kopieren
`/etc/hadoop/conf/hdfs-site.xml`
`/etc/hadoop/conf/core-site.xml`
- Zugriff des Hadoop-Clients testen
`hadoop fs -ls /user/oracle/`

SQL Connector for HDFS

- Oracle SQL Connector installieren

<https://www.oracle.com/technetwork/database/database-technologies/bdc/big-data-connectors/downloads/index.html>

```
bash> unzip oraohdfs-3.8.2.zip
```

- Umgebung anpassen

```
export JAVA_HOME=/usr
export OSCH_HOME=/home/oracle/orahdfs-3.8.2
export OSCH_BIN_PATH=$OSCH_HOME/bin
export HADOOP_CLASSPATH=$OSCH_HOME/jlib
```

- Oracle DB-Objekte anlegen:

```
# User anlegen
```

```
SQL> create user hdfuser identified by hdf;
```

```
SQL> grant create session, create table, create view to hdfuser;
```

```
SQL> grant execute on sys.utl_file to hdfuser;
```

```
# Binary access
```

```
SQL> create directory osch_bin_path as AS '/home/oracle/orahdfs-3-8-2/bin';
```

```
SQL> SQL GRANT READ, EXECUTE ON DIRECTORY OSCH_BIN_PATH TO hdfuser;
```

```
SQL> create directory external_table_dir as '/home/oracle/hdfs_tables';
```

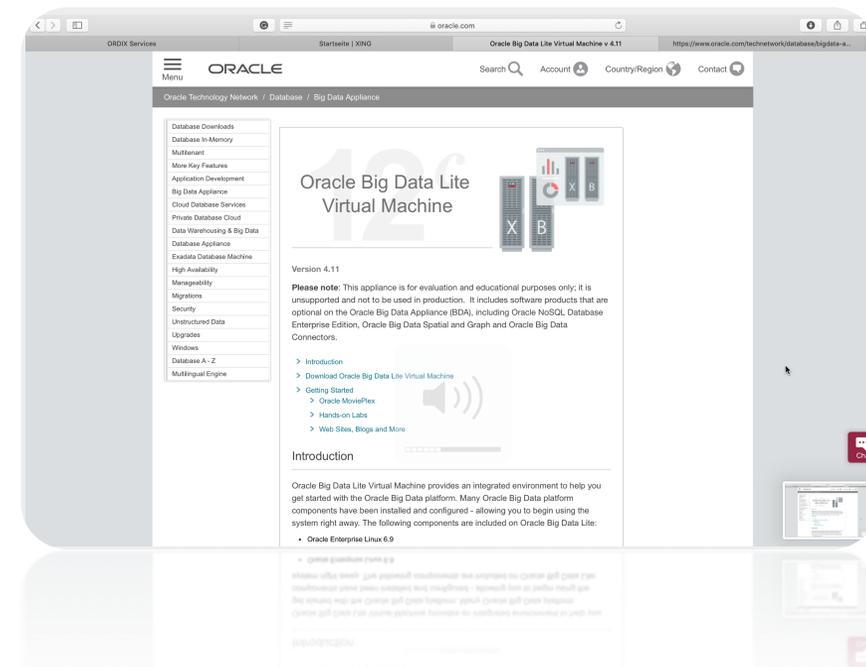
```
SQL> GRANT READ, WRITE ON DIRECTORY external_table_dir TO hdfuser;
```

- Externe Tabelle generieren (Manifest regelt die Details):

```
hadoop jar /home/oracle/orahdfs-3.8.2/jlib/orahdfs.jar
```

```
oracle.hadoop.exttab.ExternalTable -conf manifest.xml -createTable
```

- Oracle bietet eine komplette virtuelle Spielwiese an:
<https://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html>
- Voraussetzung:
 - 2 Cores
 - 5 GB RMAN
 - 50 GB Disk



ORDIX AG
Aktiengesellschaft für
Softwareentwicklung, Schulung,
Beratung und Systemintegration

Zentrale Paderborn
Karl-Schurz-Straße 19a
33100 Paderborn
Tel.: 05251 1063-0
Fax: 0180 1 67349 0

Seminarzentrum Wiesbaden
Kreuzberger Ring 13
65205 Wiesbaden
Tel.: 0611 77840-00

info@ordix.de
www.ordix.de

**Vielen Dank für
Ihre Aufmerksamkeit**