Big Data und DWH – ein Beispiel Szenario

Detlef E. Schröder
Oracle Deutschland B.V. & Co KG
Hamburg

Schlüsselworte
Big Data, Hadoop, Scenario, DWH, Oracle, Integration, Map Reduce, SQL, ETL,

Einleitung
Um das Thema Big Data ranken sich auch nach nun fast zwei Jahren Hype – Thema viele unterschiedliche Vorstellungen und Mythen. Der Vortrag will versuchen anhand eines einfachen Beispiels einiges praktisch werden zu lassen und damit zu Entmystifizieren. Dabei wird einmal der Weg von der Entstehung von Big Data bis zur Integration mit den Daten des DWH gezeigt. Damit wird dann die Möglichkeit eröffnet seine eigenen Szenarien zu entwickeln.

Architektur

Zuerst betrachten wir die Lösungsarchitektur von Oracle für das Thema Big Data und dies anhand von verschiedenen Szenarien, die das Anwendungsgebiet aus Sicht von Oracle deutlich machen.

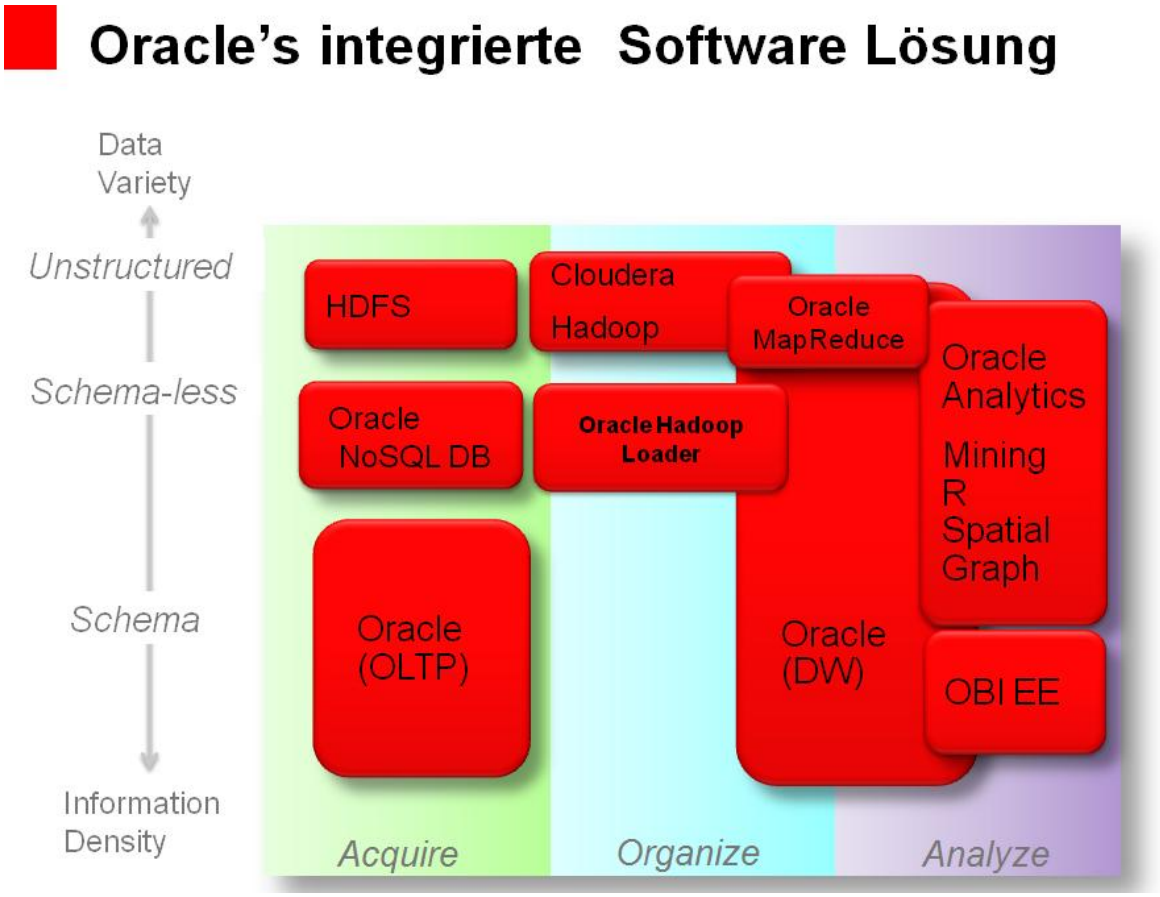

Abb. 1: Oracle Big Data Architektur

Danach beleuchten wir die einzelnen Bausteine: das Hadoop Distributed File System (HDFS), die Möglichkeiten des „Map Reduce“ und die Integration in die Oracle ETL Prozesse und Workflows.

Dies geschieht an Hand eines konkreten Beispiels.

Dazu werden Bewertungs- und Feedback-Emails von einer Website erzeugt und dann im HDFS gespeichert. Dazu werden dann auch die ersten einfachen Handgriffe mit dem HDFS gezeigt und damit die Möglichkeiten und Grenzen dargelegt.

Auszug aus der Beispiel – Datei:

```
Mail_Text:
 Produkt_Nr: 85 ->  14 Fehlerhafte_Bedienungsanleitung -> Offenkundig
super   .
Mail_Text:
 Produkt_Nr: 108 ->  14 Fehlerhafte_Bedienungsanleitung -> Das ist   eine
Zumutung   .
Mail_Text:
 Produkt_Nr: 72 -> 1 Unvollstaendig  -> Nachweisbar   prima   .
Mail_Text:
 Produkt_Nr: 38 ->  16 Falsche_Groesse -> Das war  Schrott   .
```

Anschließend wird mit Hilfe eines Map Reduce Programmes der Zugriff auf diese Daten gezeigt. Daran wird auch deutlich, wie eine Verarbeitung aussehen kann. Da die Daten ohne Verarbeitung gespeichert sind, wird die Verarbeitung hier aufgenommen.

Auszug aus dem Reducer – Programmteil zur Auswertung der Daten:

```
...
protected void reduce(Text key, Iterable<Text> values, Context context)
throws IOException, InterruptedException {
   Map<String, Integer> keyWordCounts = new HashMap<String, Integer>();
   for (Text value : values) {
      String[] split = value.toString().split("\t");
      String keyWord = split[0];
      int count = 0;
      if (keyWordCounts.containsKey(keyWord)) {
         count = keyWordCounts.get(keyWord);
      }
      count += Integer.parseInt(split[1]);
      keyWordCounts.put(keyWord, count);
   }
   String[] keySplit = key.toString().split("_");
   String product = keySplit[0];
   String failure = keySplit[1];
   for (Entry<String, Integer> entry : keyWordCounts.entrySet()) {
      String output = failure + "\t" + entry.getKey() + "\t" +
entry.getValue();
      context.write(new Text(product), new Text(output));
...
```

Bisher befanden wir uns mit dem Beispiel in der Hadoop Welt. Im Weiteren wird nun die Verbindung zur klassischen Oracle Welt aufgenommen. Dazu stellt Oracle verschiedene Wege zur Verfügung. Diese werden dargestellt und dann exemplarisch gezeigt.

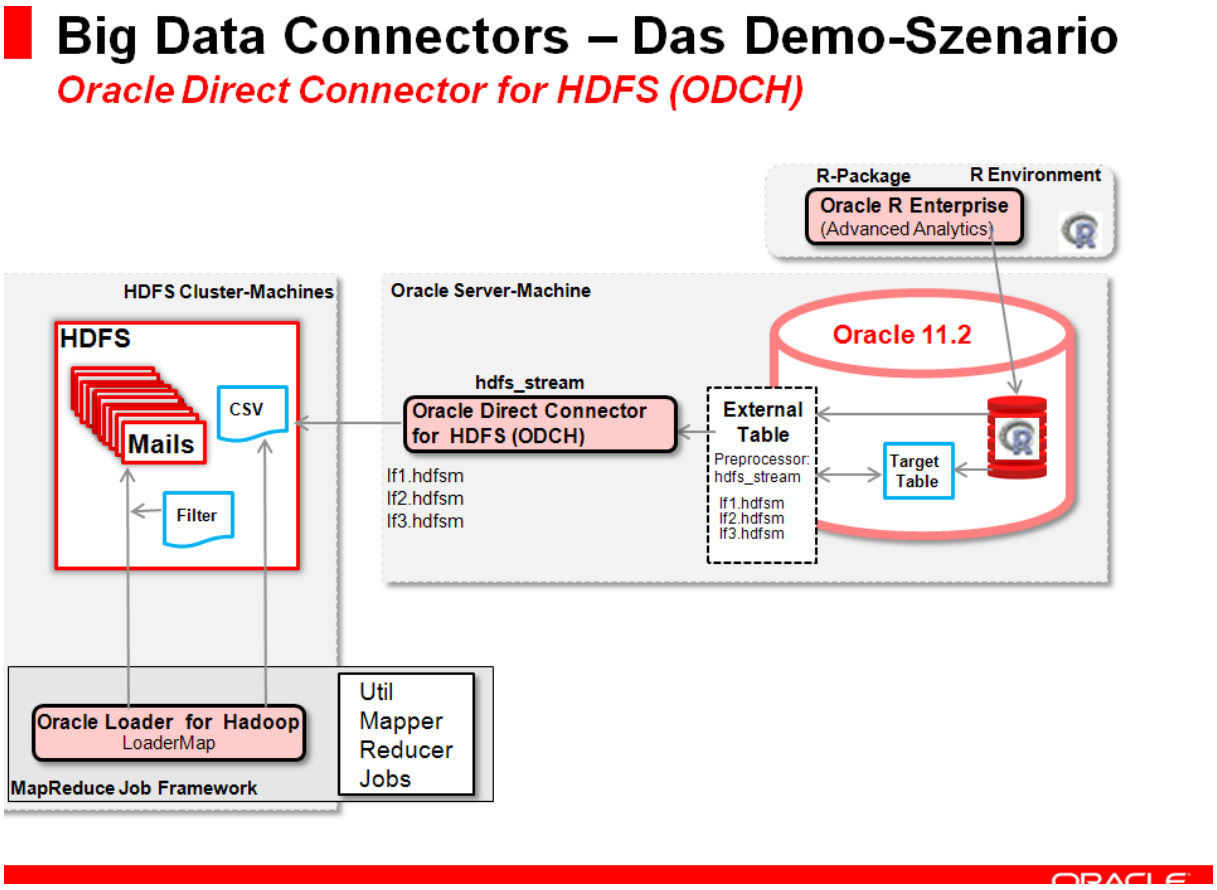

Abb. 3: Architektur des Beispieles

So lässt sich aus der Oracle Datenbank heraus eine externe Tabelle anlegen, die auf den Ergebnissen eines Map Reduce Programmes basiert. Oder aber die Ergebnisse liegen gleich als Data Pump Format vor und können dann verarbeitet werden.

```
CREATE TABLE mail_wert
(  produkt_nr NUMBER
  ,fehler_nr  NUMBER
  ,fehler_text VARCHAR2(50)
  ,wert_nr NUMBER
) ORGANIZATION EXTERNAL
(TYPE oracle_loader
 DEFAULT DIRECTORY "XTAB_DATA_DIR"
 ACCESS PARAMETERS
 (
     records delimited by newline
     preprocessor HDFS_BIN_PATH:hdfs_stream
     badfile XTAB_LOG_DIR:'fivdti_xt%a_%p.bad'
     logfile XTAB_LOG_DIR:'fivdti_xt%a_%p.log'
     fields terminated by '\t'
     missing field values are null
     (
       produkt_nr  CHAR(3),
       fehler_nr CHAR(2),
       fehler_text CHAR(50),
       wert_nr CHAR(3)
     )
 )
 LOCATION ('lf1.hdfsm', 'lf2.hdfsm', 'lf3.hdfsm')
) PARALLEL REJECT LIMIT UNLIMITED;
```

Abb. 2: Definition der Externen Tabelle

Die Verarbeitung der Inhalte der Big Data Welt bestimmt sich aber vor allem aus dem Anwendungshintergrund. Damit die notwendige Flexibilität gegeben ist, verfügt Oracle über die notwendige Schnittstellenvielfalt.

Zum Abschluss können dann die neu gewonnen Informationen zusammen mit den bestehenden aus dem DWH abgefragt werden und daraus neue Erkenntnisse gewonnen werden. Dies bildet den Abschluss des Beispiels.

Zum Abschluss werden die verschiedenen Erkenntnisse, die mit dem Beispiel gewonnen wurden noch mal zusammengefasst und auf die Anwendungsgebiete angewendet.

Kontaktadresse:
Detlef E. Schröder
Oracle Deutschland B. V. & Co KG
Kühnehöfe 5
D-22761 Hamburg

Telefon:	+49 (40) 89091-423
E-Mail	Detlef.E.Schroeder@oracle.com
Internet:	www.oracle.de