Big Data

Marriage of RDBMS-DWH and Hadoop & Co.

Author: Jan Ott – Trivadis AG

Mit über 600 IT- und Fachexperten bei Ihnen vor Ort

12 Trivadis Niederlassungen mit über 600 Mitarbeitenden

200 Service Level Agreements

Mehr als 4'000 Trainingsteilnehmer

Forschungs- und Entwicklungs-budget: CHF 5.0 Mio. / EUR 4.0 Mio.

Finanziell unabhängig und nachhaltig profitabel

Erfahrung aus mehr als 1'900 Projekten pro Jahr bei über 800 Kunden

Agenda

1. **Introduction**
2. First Steps in the Big Data – Hadoop World
3. Project 1
4. Project 2
5. Summary

Introduction

- A view words about Big Data – Hadoop
- First Steps in the Big Data – Hadoop World
  - Get some data into Hadoop
  - Get the data into Oracle
  - First performance figures
- Project 1
  - Data in Hadoop as store only
  - Transformation in DB
  - Next steps
- Project 2
  - Lambda Architecture
  - Setup
  - Next steps

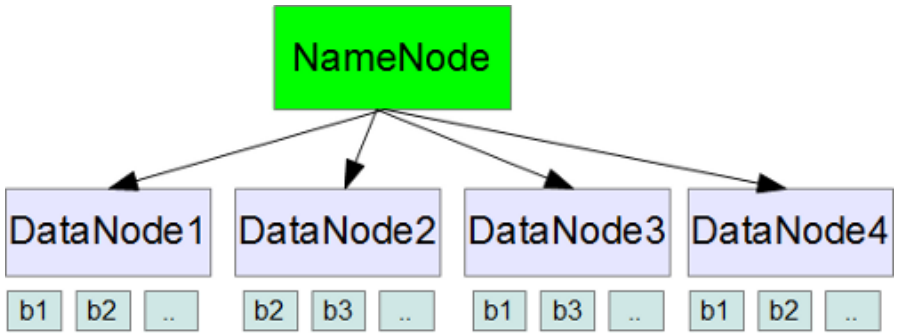
Big Data: Introduction

- Big Data - V's – 3, 4 or 5
  - Volume – scale of data
  - Velocity – analysis of streaming data
  - Variety – different form of data
  - Veracity – uncertainty of data (IBM)
  - Value – business value (Microsoft)
- Hadoop and its Zoo
  - HDFS – MapReduce
  - HBase, Hive, Impala, …
  - Zookeeper
- NoSQL Databases
- Semantic Web
- Architecture
  - LAMBDA

Turning Data into Insights

What is Hadoop

- a file system – HDFS
  - Based on papers from Google
  - Apache Open Source Project
- Goal
  - Fast
  - Handles huge amount of data
  - Handles unstructured to fully structured data
  - Horizontally scalable
  - Reliable



Agenda

1. Introduction
2. **First Steps in the Big Data – Hadoop World**
3. Project 1
4. Project 2
5. Summary

First Steps in the Big Data – Hadoop World

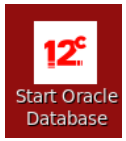
- It is a zoo and you can get lost
  - Keep it simple
- Get some data into Hadoop
- Get Oracle DB to access the data in Hadoop
- Hadoop – Java (keep it to a minimum)
- Data small and known
  - EMP table from scott/tiger
- Get an environment that is setup
  - Oracle VM – Big Data Light
- Pick one way to get the data into Hadoop
  - Hadoop shell interface
- Pick one way to make the data accessible by Oracle
  - Oracle Connectors – SQL Connector

Pre-Requisite – Environment

- Oracle Big Data Lite
  - VM
  - Version 2.4.1
  - http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html
- Contains
  - Oracle Database 12c (12.1.0.1)
  - Cloudera's Distribution including Apache Hadoop (CDH4.5)
  - Oracle Big Data Connectors 2.4
    - Oracle SQL Connector for HDFS 2.3.0
  - Oracle SQL Developer 4.0
  - …
- Oracle Virtual Box

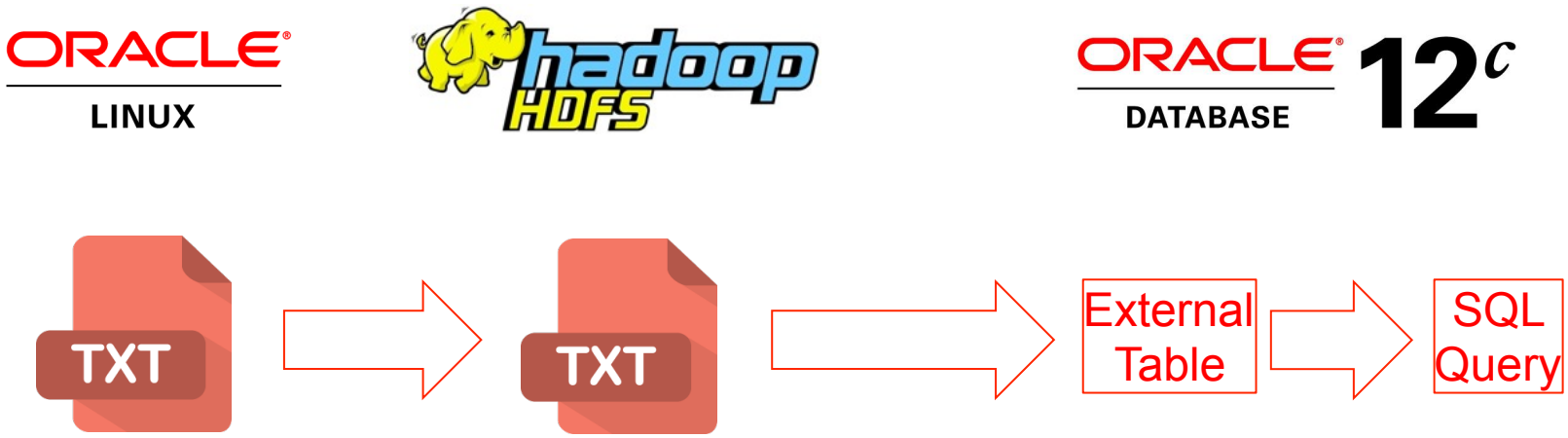
Information about the VM

- Login
  - oracle/welcome1
- Start here
  - file:///home/oracle/GettingStarted/StartHere.html
- Start the Oracle DB



- Your done preparing
- On the side – Oracle has a Movie example

The Steps – simple – focus



Step 1 – Prepare the Data

- EMP => /home/oracle/Desktop/demo/emp.txt
  - Comma delimited
  - Flat file
  - No XML – focus

```
SQL> SELECT empno || ',' || ename || ',' || job || ','
          || mgr || ',' || hiredate || ',' || sal || ','
          || comm || ',' || deptno
       FROM emp
      ORDER BY empno;
```

Step 1 – Prepare the Data

```
7369,SMITH,CLERK,7902,17-DEC-80,800,,20
7499,ALLEN,SALESMAN,7698,20-FEB-81,1600,300,30
7521,WARD,SALESMAN,7698,22-FEB-81,1250,500,30
7566,JONES,MANAGER,7839,02-APR-81,2975,,20
7654,MARTIN,SALESMAN,7698,28-SEP-81,1250,1400,30
7698,BLAKE,MANAGER,7839,01-MAY-81,2850,,30
7782,CLARK,MANAGER,7839,09-JUN-81,2450,,10
7788,SCOTT,ANALYST,7566,09-DEC-82,3000,,20
7839,KING,PRESIDENT,,17-NOV-81,5000,,10
7844,TURNER,SALESMAN,7698,08-SEP-81,1500,0,30
7876,ADAMS,CLERK,7788,12-JAN-83,1100,,20
7900,JAMES,CLERK,7698,03-DEC-81,950,,30
7902,FORD,ANALYST,7566,03-DEC-81,3000,,20
7934,MILLER,CLERK,7782,23-JAN-82,1300,,10
```

Step 2 – Get the data into HDFS

- HDFS Shell Commands
  - cat
  - chmod
  - cp
  - ls
  - put
  - …
  - http://hadoop.apache.org/docs/r2.3.0/hadoop-project-dist/hadoop-common/FileSystemShell.html
- See what there is in HDFS

```
$ hadoop fs -ls
Found 6 items
drwx------   - oracle supergroup  0 2014-03-19 11:11 .Trash
drwx------   - oracle supergroup  0 2014-01-24 20:15 .staging
drwxr-x---   - oracle supergroup  0 2014-01-13 00:15 moviedemo
drwx------   - oracle supergroup  0 2014-01-24 16:32 moviework
drwxr-xr-x   - oracle supergrou   0 2013-12-27 16:36 olhcache
drwxr-xr-x   - oracle supergroup  0 2013-12-27 16:36 temp_out_session
```

Step 2 – Get the data into HDFS

- Create a new directory

```
$ hadoop fs -mkdir demo_hdfs
$ hadoop fs -ls
Found 7 items
drwx------   - oracle supergroup  0 2014-03-19 11:11 .Trash
drwx------   - oracle supergroup  0 2014-01-24 20:15 .staging
drwxr-xr-x   - oracle supergroup  0 2014-03-19 11:21 demo_hdfs
drwxr-x---   - oracle supergroup  0 2014-01-13 00:15 moviedemo
drwx------   - oracle supergroup  0 2014-01-24 16:32 moviework
drwxr-xr-x   - oracle supergroup  0 2013-12-27 16:36 olhcache
drwxr-xr-x   - oracle supergroup  0 2013-12-27 16:36 temp_out_session
```

- Copy the file into it

```
$ hadoop fs -put /home/oracle/Desktop/demo/emp.txt demo_hdfs
$ hadoop fs -ls -R demo
-rw-r--r-- 1 oracle super… 1218 2014-03-19 11:30 demo_hdfs/emp.txt
```

Step 2 – Get the data into HDFS

- See what is there

```
$ hadoop fs -cat demo_hdfs/emp.txt
7369,SMITH,CLERK,7902,17-DEC-80,880,,20
7499,ALLEN,SALESMAN,7698,20-FEB-81,1600,300,30
7521,WARD,SALESMAN,7698,22-FEB-81,1250,500,30
7566,JONES,MANAGER,7839,02-APR-81,2975,,20
7654,MARTIN,SALESMAN,7698,28-SEP-81,1250,1400,30
7698,BLAKE,MANAGER,7839,01-MAY-81,2850,,30
7782,CLARK,MANAGER,7839,09-JUN-81,2450,,10
7788,SCOTT,ANALYST,7566,19-APR-87,3000,,20
7839,KING,PRESIDENT,,17-NOV-81,5000,,10
7844,TURNER,SALESMAN,7698,08-SEP-81,1500,0,30
7876,ADAMS,CLERK,7788,23-MAY-87,1100,,20
7900,JAMES,CLERK,7698,03-DEC-81,950,,30
7902,FORD,ANALYST,7566,03-DEC-81,3000,,20
7934,MILLER,CLERK,7782,23-JAN-82,1300,,10
```

- That's it – the data is in Hadoop – HDFS

Oracle SQL Connector

- Connects HDFS with an Oracle External Table
- Sources can be
  - Data Pump File
  - Hive Table
  - Delimited Text File
- Uses Oracle Loader
- Parallel – needs several files
- Limits
  - Read only
  - No Indexes
  - Always full table scan
- Location File
- Oracle provides – command line tool
  - http://docs.oracle.com/cd/E37231_01/doc.20/e36961/sqlch.htm

Step 3 – Setup External Table to HDFS

- Rights
  - OSCH_BIN_PATH
  - Default Directory
    - Log Files
    - Location Files

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

Grant succeeded.

SQL> CREATE OR REPLACE DIRECTORY demo_dir
        AS '/home/oracle/Desktop/demo';

Directory created.

SQL> GRANT READ, WRITE, EXECUTE ON DIRECTORY demo_dir TO scott;

Grant succeeded.
```

Step 3 – Setup External Table to HDFS

- Create External Table – Command Line Tool


```
$ hadoop oracle.hadoop.exttab.ExternalTable \
-D oracle.hadoop.exttab.tableName=EMP_HDFS_EXT_TAB \
-D
oracle.hadoop.exttab.columnNames=empno,ename,job,mgr,hiredate,sal,comm,
deptno \
-D oracle.hadoop.exttab.locationFileCount=1 \
-D oracle.hadoop.exttab.dataPaths=demo_hdfs/*.txt \
-D oracle.hadoop.exttab.columnCount=8 \
-D oracle.hadoop.exttab.defaultDirectory=demo_dir \
-D oracle.hadoop.connection.url=jdbc:oracle:thin:@localhost:1521:orcl \
-D oracle.hadoop.connection.user=SCOTT \
-D oracle.hadoop.exttab.printStackTrace=true \
-createTable
```

Step 3 – Setup External Table to HDFS


```
CREATE TABLE "SCOTT"."EMP_HDFS_EXT_TAB"
(
 "EMPNO"                                 VARCHAR2(4000),
 "ENAME"                                 VARCHAR2(4000),
 "JOB"                                   VARCHAR2(4000),
 "MGR"                                   VARCHAR2(4000),
 "HIREDATE"                              VARCHAR2(4000),
 "SAL"                                   VARCHAR2(4000),
 "COMM"                                  VARCHAR2(4000),
 "DEPTNO"                                VARCHAR2(4000)
)
ORGANIZATION EXTERNAL
(
   TYPE ORACLE_LOADER
   DEFAULT DIRECTORY "DEMO_DIR"
   ACCESS PARAMETERS
   (
     RECORDS DELIMITED BY 0X'0A'
     CHARACTERSET AL32UTF8
     PREPROCESSOR "OSCH_BIN_PATH":'hdfs_stream'
     FIELDS TERMINATED BY 0X'2C'
```

Step 3 – Setup External Table to HDFS


```
...(
     RECORDS DELIMITED BY 0X'0A'
     CHARACTERSET AL32UTF8
     PREPROCESSOR "OSCH_BIN_PATH":'hdfs_stream'
     FIELDS TERMINATED BY 0X'2C'
     MISSING FIELD VALUES ARE NULL
     (
       "EMPNO" CHAR(4000),
       "ENAME" CHAR(4000),
       "JOB" CHAR(4000),
       "MGR" CHAR(4000),
       "HIREDATE" CHAR(4000),
       "SAL" CHAR(4000),
       "COMM" CHAR(4000),
       "DEPTNO" CHAR(4000)
     )
   )
   LOCATION
   (
     'osch-20140319061232-9144-1'
   )
) PARALLEL REJECT LIMIT UNLIMITED;
```

Step 3 – Setup External Table to HDFS

- Creates a location file
  - Filename Pattern: osch-timestamp-number-n

```
cat /home/oracle/Desktop/demo/osch*
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<locationFile>
    <header>
        <version>1.0</version>
        <fileName>osch-20140319061232-9144-1</fileName>
        <createDate>2014-03-19T18:12:32</createDate>
        <publishDate>2014-03-19T06:12:32</publishDate>
        <productName>Oracle SQL Connector for HDFS Release 2.3.0 -
Production</productName>
        <productVersion>2.3.0</productVersion>
    </header>
    <uri_list>
        <uri_list_item size="605" compressionCodec="">hdfs://
bigdatalite.localdomain:8020/user/oracle/demo_hdfs/emp.txt</
uri_list_item>
    </uri_list>
</locationFile>
```

Step 3 – Setup External Table to HDFS

- Creates log/bad files

```
cat /home/oracle/Desktop/demo/EMP_HDFS_EXT_TAB*
 LOG file opened at 03/19/14 20:06:31

KUP-05007:   Warning: Intra source concurrency disabled because the
preprocessor option is being used.

Field Definitions for table EMP_HDFS_EXT_TAB
  Record format DELIMITED, delimited by 0A
  Data in file has same endianness as the platform
  Rows with all null fields are accepted

  Fields in Data Source:

    EMPNO                               CHAR (4000)
      Terminated by "2C"
      Trim whitespace same as SQL Loader
    ENAME                               CHAR (4000)
      Terminated by "2C"
      Trim whitespace same as SQL Loader
...
```

Performance

- It is for HUGE amounts of data
  - Small files are just slow

# of Files	Size per File	Total Size	# of Rows per File	Total # of Rows	Time
1	605 B	605 B	14	14	00:53.17
10	605 B	6 KB	14	140	00:52.06
10	422 KB	4.2 MB	10’000	100’000	00:51.68
10	42 MB	420 MB	1’000’000	10’000’000	01:00.92
50	42 MB	2.1 GB	1’000’000	50’000’000	01:27.81

Agenda

1. Introduction
2. First Steps in the Big Data – Hadoop World
3. **Project 1**
4. Project 2
5. Summary

Project 1 – Pilot Phase

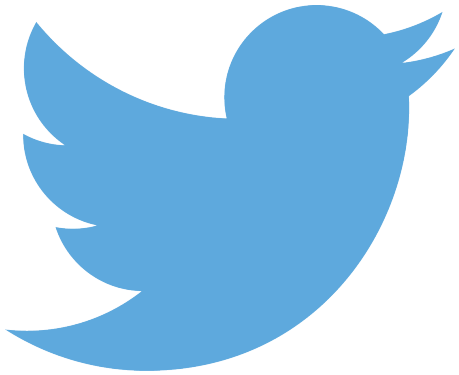
- Move to Hadoop for delivered files
  - Start collecting
  - Files get copied into HDFS one to one
  - No decision had to be taken
    - Schema – schema-less
    - Table design - non
    - …
  - Immutable Data Store - Create and Read
    - No update / No delete
- Add External Tables with ORACLE SQL Connector
  - Data useable
- Build Hadoop infrastructure
- Next
  - Analyze the Zoo

Agenda

1. Introduction
2. First Steps in the Big Data – Hadoop World
3. Project 1
4. **Project 2**
5. Summary

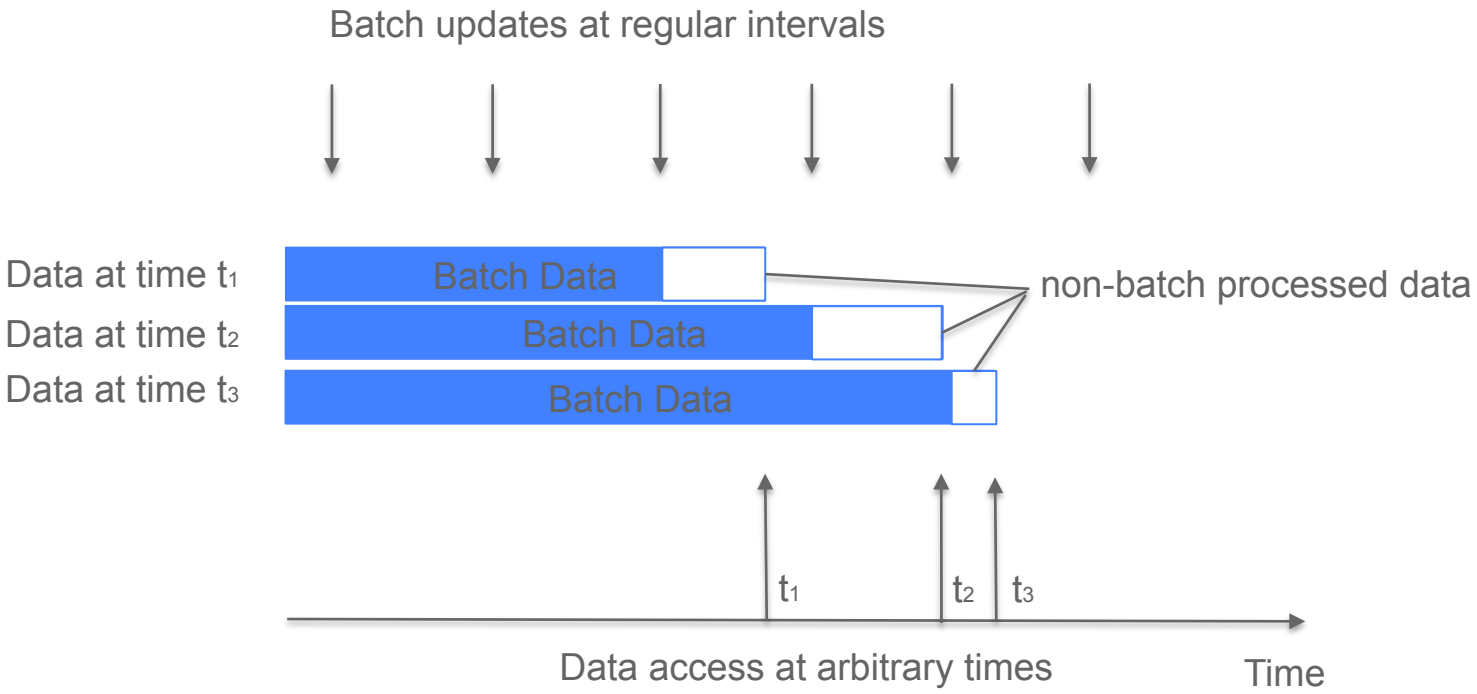
Project 2 – Figures

- 400 – 500 Mio tweets per day
- 1 tweet contains
  - Around 50 metadata pieces
    - Geo-location
    - Re-tweets
    - Followers
  - That is about 2 A4 pages
- Twitter Sample Stream
  - 1%
  - 4-5 Mio tweets per day
  - 50 tweets per second
- 20 other streams with defined key words
- HDFS
  - 1 TB every 2 months including replication

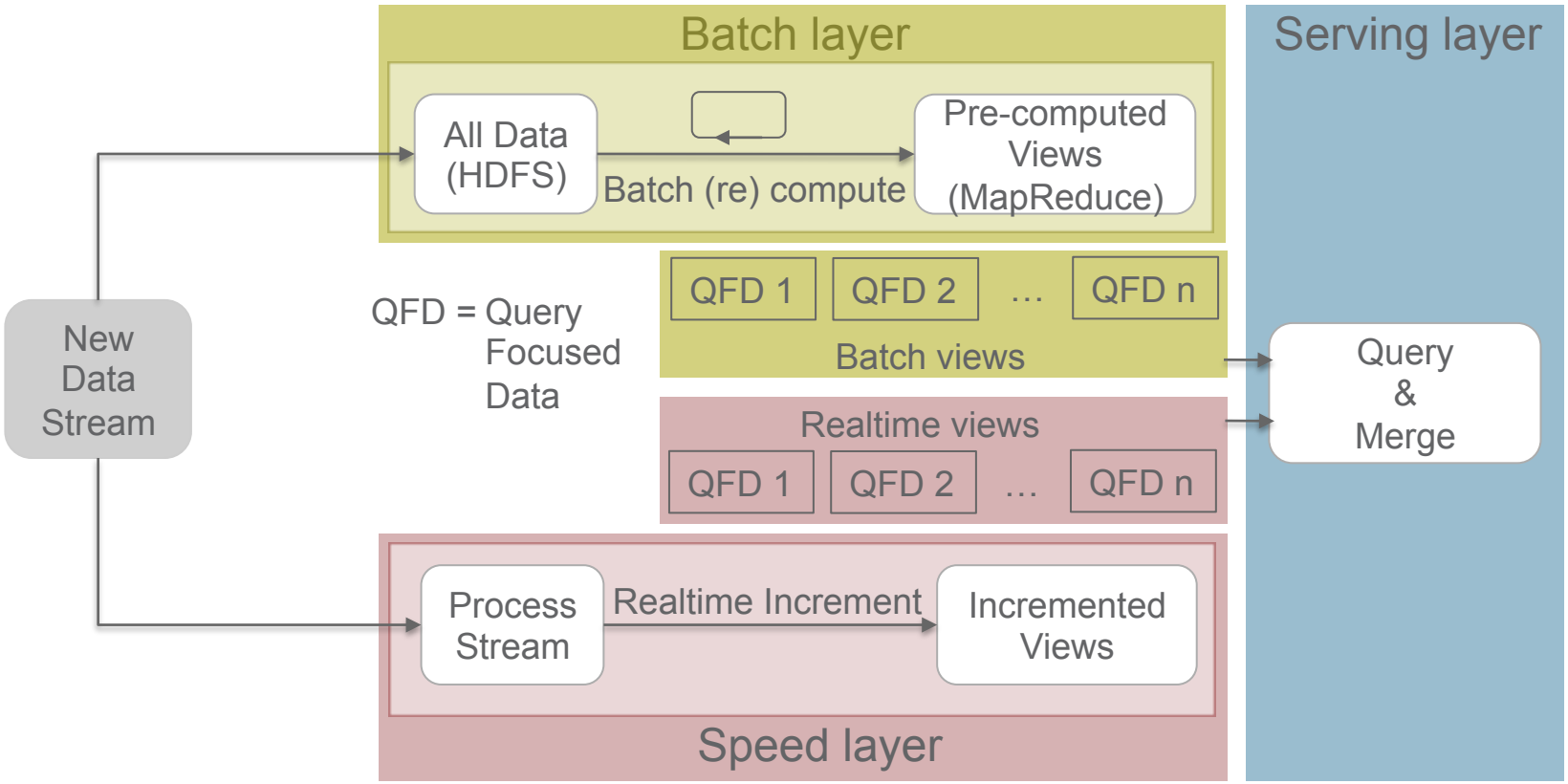


Problem – history data and actual data

- The problem with simple batch approaches: missing data!

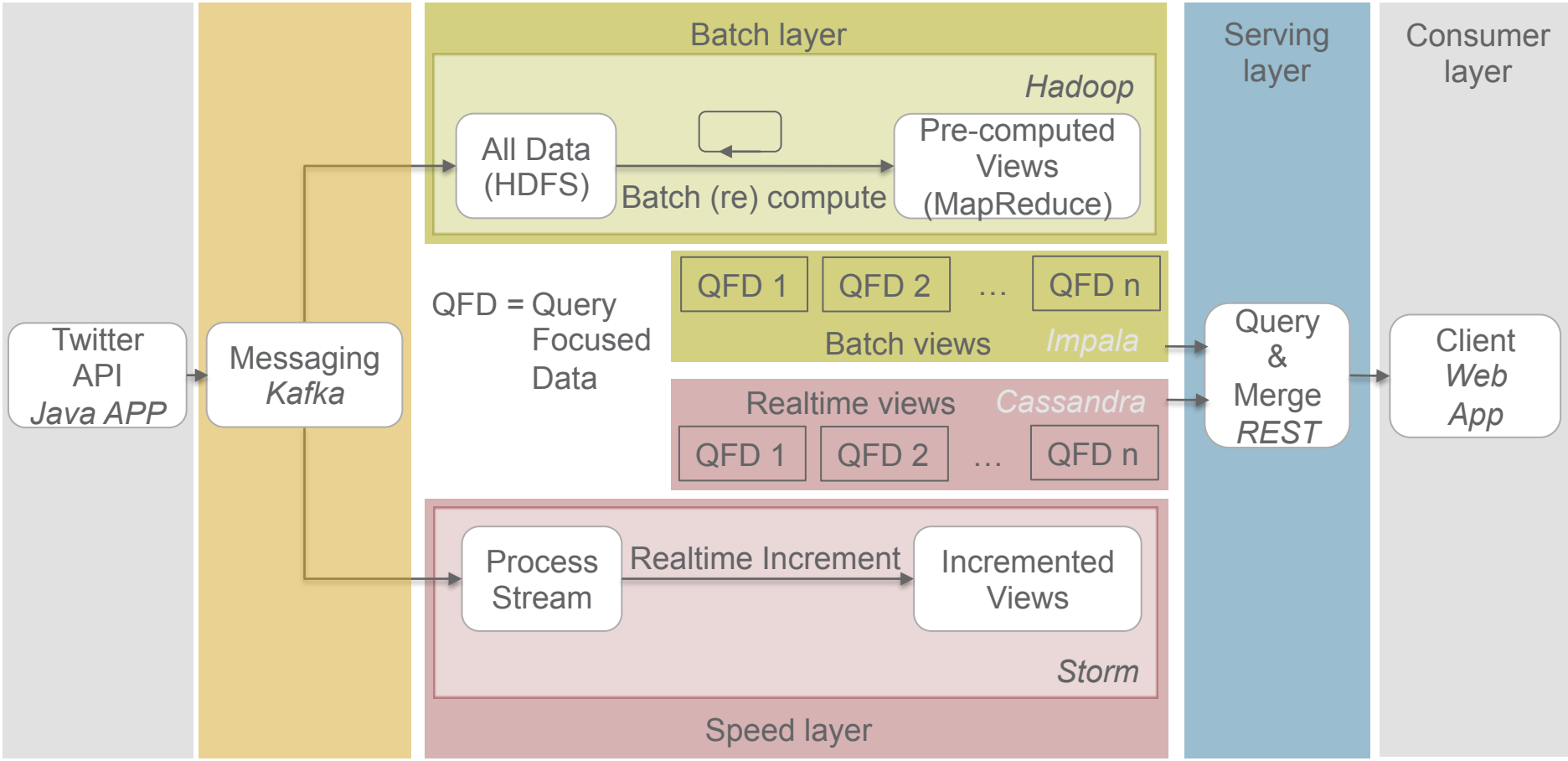


The Lambda Architecture



adapted from Kinley (2013)

The Lambda Architecture - adopted



Project 2

- Live streams
- Batch computation
- Scalable
- Open for new sources

Agenda

1. Introduction
2. First Steps in the Big Data – Hadoop World
3. Project 1
4. Project 2
5. **Summary**

Summary

- Big Data <> Hadoop
- Hadoop – HDFS
  - File System
  - Block Size – 128 MB
  - Nothing for small files
  - No optimization with indexes
- A new World
- Hadoop and its Zoo
- Lots can be done with RDBMS
- Start to collect now

Questions?

THANK YOU.

Trivadis AG

Jan Ott

Europa-Strasse 5
CH-8152 Glattbrugg-Zurich

Tel. +41-44-808 70 20 (reception)
Fax +41-44-808 70 21

info@trivadis.com
www.trivadis.com

BASEL BERN LAUSANNE ZÜRICH DÜSSELDORF FRANKFURT A.M. FREIBURG I.BR. HAMBURG MÜNCHEN STUTTGART WIEN

Sources

- Oracle Connection to Hadoop with Oracle
  - https://blogs.oracle.com/bigdataconnectors/entry/how_to_load_oracle_tables
- Pictures
  - Oracle.com
  - Twitter.com
  - Apache.com
- Big Data – MEAP by Nathan Marz