# Hive über HBase

Als erstes laden wir Traffic-Daten von Wikipedia herunter. Darin enthalten sind Sprache, Titel der Seite, Anzahl der Zugriffe und die Größe in Bytes.

mkdir pagecounts ; cd pagecounts

for x in {0..9} ; do wget "http://dumps.wikimedia.org/other/pagecounts-raw/2008/2008-10/pagecounts-20081001-0${x}0000.gz" ; done

hadoop fs -copyFromLocal $(pwd) ./

Anschließend erstellen wir eine Hive Tabelle aus den Daten (in hive-shell):

CREATE TABLE IF NOT EXISTS pagecounts (projectcode STRING, pagename STRING, pageviews STRING, bytes STRING)

ROW FORMAT

DELIMITED FIELDS TERMINATED BY ' '

LINES TERMINATED BY '\n'

STORED AS TEXTFILE

LOCATION '/user/cloudera/pagecounts';

Schema für HBase transformieren. Rowkey ist konkatenierte Projektcode, Seitenname und das Datum mittels / getrennt (in hive-shell):

CREATE VIEW IF NOT EXISTS pgc (rowkey, pageviews, bytes) AS

SELECT concat\_ws('/',

projectcode,

concat\_ws('/',

pagename,

regexp\_extract(INPUT\_\_FILE\_\_NAME, 'pagecounts-(\\d{8}-\\d{6})\\..\*$', 1))),

pageviews, bytes

FROM pagecounts;

SELECT \* FROM pgc WHERE rowkey LIKE 'en/q%' LIMIT 10;

Nun wird die Tabelle in HBase registiert (in hive-shell):

CREATE TABLE IF NOT EXISTS pagecounts\_hbase (rowkey STRING, pageviews STRING, bytes STRING)

STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'

WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,f:c1,f:c2')

TBLPROPERTIES ('hbase.table.name' = 'pagecounts');

Damit die HBase library genutzt werden kann, muss der CLASSPATH angepasst werden:

export HADOOP\_CLASSPATH=/etc/hbase/conf:/usr/lib/hbase/hbase-0.94.6.1.3.2.0-111-security.jar:/usr/lib/zookeeper/zookeeper.jar

Daten ins HBase laden(in hive-shell):

SET hive.aux.jars.path = file:///etc/hbase/conf/hbase-site.xml,file:///usr/jars/hbase-server-1.0.0-cdh5.5.0.jar,file:///usr/jars/zookeeper-3.4.5-cdh5.5.0.jar;

FROM pgc INSERT INTO TABLE pagecounts\_hbase SELECT pgc.\* WHERE rowkey LIKE 'en/q%' LIMIT 10;

Daten anzeigen (in HBase-shell):

scan 'pagecounts'

Daten mit Hive vergleichen (in hive-shell):

SELECT \* from pagecounts\_hbase;

Daten von hive analysieren (in hive-shell):

SET hive.aux.jars.path = file:///etc/hbase/conf/hbase-site.xml,file:///usr/jars/hbase-server-1.0.0-cdh5.5.0.jar,file:///usr/jars/zookeeper-3.4.5-cdh5.5.0.jar;

SELECT count(\*) from pagecounts\_hbase;