Hive über HBase

Bei diesem Beispiel werden Traffic-Statistiken von Wikipedia als Beispieldaten genutzt.

1. **Daten herunterladen**
   - `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) ./`

2. **Hive Tabelle erstellen**
   - ```
     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';
     ```

3. **Rohdaten für HBase Schema transformieren**
   - ```
     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;`

Der RowKey ist eine Zusammensetzung aus `projectcode`, `pagename`, und `date` getrennt durch /.

4. **Tabelle in HBase registrieren**
   - ```
     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');
     ```

5. **Classpath anpassen (zur Nutzung der HBase library)**
   - ```
     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
     ```

6. **Daten in HBase laden (via 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;`

7. **Daten Anzeigen**
   a. **via HBase shell**
      - `scan 'pagecounts'`

   b. **via Hive shell**
      - `SELECT * from pagecounts_hbase;`

8. **Analyse durch Hive (via 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;`