Björn Lohrmann
bjoern.lohrmann@tu-berlin.de
FG CIT

Cloud Computing
Summer Term 2012

Project Assignment No. 4

DUE: 12.07.2012 23:59

The primary goal of this exercise is to gain insight in comparing the Hadoop and Stratosphere frameworks for parallel data processing.

1. Prerequisites

- TPC-H is a benchmark suite for relational databases and provides a data model (entities and their relationships) as well as a set of queries. The TPC-H data model describes a fictional enterprise with customers that have orders which consist of lineitems. In this assignment, you will only need TPC-H to generate test data for your applications. The TPC-H tools are freely available here and documentation can be found here. The TPC-H tool `dbgen` generates pipe-delimited ASCII files, one for each entity.
- Please use the following framework versions to write your code:
  - v1.0.3 of the Apache Hadoop framework
  - v0.1.2 of the Stratosphere framework
- Documentation and examples on how to use the frameworks is available on the Hadoop and Stratosphere websites.

2. Using Hadoop to determine „large volume customers"

Using the TPC-H tools, generate a data-set with the (default) scaling factor 1. Make yourself familiar with the `customer`, `order` and `lineitem` entities and their relationships (see section 1.2 of TPC-H documentation), as these will be required for this exercise.

Based on the generated TPC-H data, write a MapReduce program using Hadoop 1.0.3 that finds all orders with a volume larger than X, where X can be supplied as a command-line parameter. The volume of an `order` is the sum over the `quantity` values of all `lineitems` in the `order`.

For each order fulfilling that condition, your program shall output the name and key of the customer who submitted the order, as well as the order's key and the total quantity of lineitems in the order.

In your submission, provide a schematic overview of your MapReduce graph and your Hadoop job's source code. To complete this assignment at least on member of your group also has to give a live demonstration of your Hadoop job(s).

Hints:

- TPC-H is a benchmark for relational databases, but you must not use any SQL or other database to complete this exercise. Use Hadoop to read and process the raw ASCII data in the pipe-delimited `.tbl` files and the process the query.

- You may use two ore more successive MapReduce jobs to compute the final query result.
- Joining two relations with MapReduce is tricky but possible. As this is a common problem some research on the Internet may be helpful if you cannot come up with a solution of your own.

3. Using Stratosphere to determine „large volume customers"

Use PACTs to compute the same query as before. In your submission, provide a schematic overview of your PACTs graph and your PACT job's source code. To complete this assignment at least on member of your group also has to give a live demonstration of your PACT job.

Hints:

- Joining two relations is easier in PACTs than in MapReduce, hence you can (and must) not only use map and reduce PACTs.
- You can use the `cc-uebung4` folder from the additional material file to get you started. It is recommended to use Maven to build the jar files required for running PACT jobs. If you change the PACT PlanAssembler class (`CustomerSummaryJob` in the tutorial example), you have to update the „`Pact-Assembler-Class`" manifest setting in Maven's `pom.xml`. You can build the example by running „`mvn package`" on the shell.

4. Submission Deliverables

Your submission on ISIS should be a single .zip file containing

- The schematic overview of your MapReduce graph and the source code from section 2.
- The schematic overview of your PACTs graph and the source code from section 3.
- A file called „group.txt" that contains the full names of all group members including their matriculation numbers.

To complete this assignment at least one member of your group has to give a live demonstration of your applications at 13.07.2012 12-14 in EN057.