# Top K Patent Citations

In diesem fortgeschrittenen MapReduce-Beispiel geht es darum, die Top K Einträge einer Liste zu ermitteln. In diesem Beispiel wird eine gekürzte Version der Patentliste cite75\_99.txt als Datenquelle verwendet. (Gekürzt auf 100.000 Zeilen, um die Größe des Abgabe-Repositories nicht unnötig zu erhöhen).

In diesem Beispiel wird gezählt, wie oft ein Patent von anderen Patenten zitiert wurde. Anschließend werden die Top 10 Patente in einer Liste im HDFS ausgegeben. Dazu wurden 2 MapReduce Jobs hintereinander geschaltet.

Die Datei ist im Format: „CITING\_PATENT“, “CITED\_PATENT“

Der erste Mapper extrahiert den „CITED“-Part und schreibt ein Tupel (CITED, 1). Der Reducer aggregiert dies zu (CITED, n). Die Tupel werden dann in eine PriorityQueue mit eigenem Comparator gespeichert. Dadurch werden die Einträge der Queue absteigend sortiert und sobald die Liste länger als K wird, wird das letzte Item (mit dem geringsten Wert) entfernt.

Der zweite Mapper liest die Daten des ersten Jobs und schreibt die Tupel (CITED, n) an einen einzelnen Reducer (job2.setNumReduceTasks(1);) welcher die Daten in einer PriorityQueue aggregiert, um die Top K Einträge über alle Splits zu produzieren.

Beide Jobs nutzen dabei dieselbe Reducer Implementierung.

unzip cite.zip

hadoop fs –put cite\_small.txt cite\_small.txt

hadoop jar topk.jar cite\_small.txt topcite

hadoop fs -cat topcite/part-r-00000

* 3658634 10
* 3538166 10
* 3385819 8
* 3470943 8
* 3560441 8
* 3578819 8
* 3706549 8
* 3717609 8
* 3773558 7
* 3697507 7