HACKATHON

Max Burgert, Tom Cinbis, Paul Höft und Tim Königl

8. Mai 2017

1 Aufgabe

Die Aufgabenstellung für diesen Hackathon lautete, aus ca. zwei Milliarden URLS auf einem Laptop in Echtzeit fünf Vorschläge für eine Auto-Completion zu liefern und ausserdem die Anzahl der möglichen Treffer anzuzeigen.

2 Ansatz

Da die Datenmenge in einer 115 GB grossen .txt Datei gespeichert war, bedurfte es eines Prozesses, diese Datenmenge erst vorzubereiten um anschliessend eine Auto-Completion darauf in angemessener Zeit ausführen zu können.
Um mit dieser Datenmenge arbeiten zu können, entschieden wir uns für die Nutzung einer Datenbank. Besonders geeignet hierfür schien uns PostrgeSQL, da diese Datenbank eine optimierte Einlesefunktion grosser Datenmengen in eine Datenbank umfasst.
Als Programmiersprache entschieden wir uns für Java, in welcher wir uns am Besten auskannten.

3 Optimierung

Eine Indizierung dieser grossen Datenmenge erwies sich in PostgreSQL als unvorteilhaft, da die standardmässige B-tree Indizierung die maximale Bytegrösse überschritten hat und ein Hashen dieser unbrauchbar war.
Um besonders die Berechnung für die Anzahl der Treffer deutlich zu beschleunigen, teilten wir die ursprüngliche Datenbank, welche die gesamte Datenmenge umfasste, anhand von „http“, „https“und anschliessend nach „www“, sowie nach den Anfangsbuchstaben „a-z“auf.
Bei entsprechender Eingabe musste so nicht mehr über die gesamte Datenmenge nach passenden Einträgen gesucht werden, sondern nur in der jeweiligen Datenbank.
Für Eingaben, welche Fälle beeinhalten bei dem das Aufteilen der Datenbanken keinen Laufzeitvorteil erbrachten, haben wir die Berechnung der Anzahl der Treffer vorberechnet.