Apache Hadoop Installation

Für die Installation von Hadoop verwenden wir die Cloudera Distribution.

Dazu fügen wie zuerst die Paketquellen für die entsprechende Version hinzu. Zum Zeitpunkt des Wiki Beitrags ist die CDH3 aktuell. Ihr könnt das Beispiel natürlich auch auf eine andere Version anwenden:

DISTRO müsst ihr durch eure verwendete Version ersetzen. Verwendet ihr Ubuntu 10.10, muss es also maverick lauten.

Quellcode

1. sudo bash -c 'cat < /etc/apt/sources.list.d/cloudera.list
2. deb http://archive.cloudera.com/debian DISTRO-cdh3 contrib
3. deb-src http://archive.cloudera.com/debian DISTRO-cdh3 contrib
4. EOF'

Anschließend signieren wir den APT Key von Cloudera:

Quellcode

1. wget -q -O - http://archive.cloudera.com/debian/archive.key | sudo apt-key add -

Danach können wir die Paketlisten aktualisieren

Quellcode

1. sudo apt-get update

Danach installieren wir das Hadoop System mit

Quellcode

1. sudo apt-get install hadoop-0.20

Nun müssen wir uns überlegen welche Rollen der Hadoop Node denn übernehmen soll. Im Beispiel entscheiden wir uns für ein Single Node Setup. Das heißt, dass alle Hadoop Komponenten erstmal auf einem Computer laufen sollen. Dazu installieren wir die folgenden Init Skripte.

Quellcode

1. sudo apt-get install hadoop-0.20-namenode hadoop-0.20-datanode hadoop-0.20-secondarynamenode hadoop-0.20-jobtracker hadoop-0.20-tasktracker

Wollt ihr Hadoop auf nur einem Node betreiben, dann könnt ihr euch die Pseudo Konfiguration wie folgt installieren:

Quellcode

1. sudo apt-get install hadoop-0.20-conf-pseudo

== Hadoop als Entwicklungssystem ==
Wenn ihr euren Rechner eher selten für Hadoop nutzt, dann könnt ihr verhindern, dass Hadoop bei jedem Hochfahren des

Rechners startet indem ihr die Initscripte aus den Runlevels entfernt:

Quellcode

```
1. for i in /etc/init.d/hadoop-*; do sudo update-rc.d -f `basename $i` remove; done
```

Um alle Services zu starten oder stoppen könnt ihr dann z.B. folgenden Code nutzen:

Quellcode

```
1. for service in /etc/init.d/hadoop-0.20-*; do sudo $service restart; done
```

== Hadoop Frontend ==
Hue ist ein grafisches Frontend für Hadoop. Gerade wenn ihr mit Hadoop beginnt, macht es Spaß damit zu arbeiten, ihr könnt einfach das HDFS browsen und den Status von Map/Reduce Tasks beobachten.

Ihr könnt Hue aus der Cloudera Distribution installieren:

Quellcode

```
1. sudo apt-get install hue
```

Wollt ihr verhindern, dass hue beim Computerstart geladen wird, könnt ihr die Runlevel Links wie folgt entfernen:

Quellcode

```
1. sudo update-rc.d -f hue remove
```

Danach könnt ihr über euren Browser auf Hue zugreifen: localhost:8088/

== Literatur ==

- wiki.cloudera.com/display/DOC/CDH3+Installation