Hadoop mit LZO

Der Lempel-Ziv-Oberhumer (LZO) Datenkompressionsalgorithmus ist ein verlustfreier Datenkompressionsalgorithmus, der eine hohe Geschwindigkeit beim Entpacken erreicht. Das Hadoop LZO Projekt bietet paralleles arbeiten mit gesplitteten LZO Dateien.

Wir verwenden die Hadoop LZO Distribution wie sie bei Twitter eingesetzt wird.
Voraussetzung für diese Anleitung ist, dass ihr Hadoop bereits wie hier geschildert installiert habt: [wiki]Apache Hadoop Installation[/wiki]
Diese baut auf der Hadoop GPL Compression auf, enthält aber einige Verbesserungen die noch nicht in das andere Projekt übernommen wurden.

Zuerst beginnen wir damit LZO zu installieren. Auf einem Debian basierten System installiert ihr lzo über
Quellcode

1. sudo apt-get install liblzo2-dev

Wir laden nun die aktuellen Dateien von Hadoop-LZO aus den Quellen bei github herunter und entpacken diese
Quellcode

1. wget --no-check-certificate https://github.com/kevinweil/hadoop-lzo/tarball/master
2. tar xvfz master

Für das Kompilieren der Native Extensions muss $JAVA_HOME korrekt gesetzt sein, mit echo können wir den aktuellen Wert prüfen und ihn dann gegebenenfalls überschreiben:
Quellcode

1. echo $JAVA_HOME
2. export JAVA_HOME=/usr/lib/jvm/java-6-sun

Dann wechseln wir in das Verzeichnis und kompilieren die Native Extension
Quellcode

1. cd kevinweil-hadoop-lzo*
2. ant compile-native tar

Nach einem erfolgreichen Build befinden sich die notwendigen Dateien im build Ordner. Sowohl die JAR Datei, als auch die Java Native Extensions müssen nun in Hadoop platziert werden:
Quellcode

1. # Copy the jar file
2. sudo cp build/hadoop-lzo-*.jar /usr/lib/hadoop-0.20/lib/
3. # Copy the native library
4. sudo tar -cBf - -C build/hadoop-lzo-*/lib/native . | sudo tar -xBvf - -C /usr/lib/hadoop-0.20/lib/native

Außerdem muss LZO in der Konfiguration aktiviert werden. Dazu legen wir eine neue Konfigurationsdatei an, falls diese noch nicht existiert.
Sie muss die zwei nachfolgenden Einstellungen enthalten:

sudo vim /etc/hadoop/conf/core-site.xml

Quellcode

```
<property>
<name>io.compression.codecs</name>
<value>org.apache.hadoop.io.compress.GzipCodec,org.apache.hadoop.io.compress.DefaultCodec,com.hadoop.compressio
</property>
<property>
<name>io.compression.codec.lzo.class</name>
<value>com.hadoop.compression.lzo.LzoCodec</value>
</property>
```

Nachdem alle Änderungen vorgenommen wurden, muss Hadoop noch neugestartet werden. Je nachdem welche Services ihr laufen habt, könnt ihr das entweder manuell machen oder mit folgendem Aufruf.

Quellcode

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

== Funktionstest ==
Nun könnt ihr einen Hadoop LZO einem einfachen Funktionstest unterziehen. Erstellt eine LZO Index Datei.
Dazu müsst ihr erstmal auf eurem Desktopcomputer eine LZO Datei erstellen können, installiert euch dazu das Tool: lzop

Quellcode

```
sudo apt-get install lzop
```

Nun sucht euch eine große Datei aus, die ihr indexieren wollt und macht eine LZO Datei daraus. Im Beispiel entscheide ich mich für big_file.txt.
Diese müsst ihr anschließend in das HDFS kopieren.

Quellcode

```
# compress any file
lzop big_file.txt
# upload to hadoop
hadoop dfs -put big_file.txt.lzo /tmp/
```

Nun indexiert die Datei:

Quellcode

```
hadoop jar /usr/lib/hadoop-0.20/lib/hadoop-lzo-*.jar com.hadoop.compression.lzo.LzoIndexer /tmp/big_file.txt.lzo
```

Quellcode

```
hadoop jar /usr/lib/hadoop-0.20/lib/hadoop-lzo-*.jar com.hadoop.compression.lzo.LzoIndexer /tmp/big_file.txt.lzo
INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev fatal: Not a git repository (or any of the parent directories): .git]
INFO lzo.LzoIndexer: [INDEX] LZO Indexing file /tmp/big_file.txt.lzo, size 0,00 GB...
INFO lzo.LzoIndexer: Completed LZO Indexing in 0,14 seconds (0,29 MB/s). Index size is 0,01 KB.
```

Und schon sollte eine LZO Index Datei entstanden sein:

Quellcode

```
1. hadoop dfs -ls /tmp/
```

Quellcode

```
1. Found 2 items
2. -rw-r--r-- 1 tb supergroup 42456 2011-03-13 11:05 /tmp/big_file.txt.lzo
3. -rw-r--r-- 1 tb supergroup 8 2011-03-13 11:05 /tmp/big_file.txt.lzo.index
```