Universität Hamburg
Department Informatik
Wissenschaftliches Rechnen
Dr. J. Kunkel, H. Hatzel, J. Gresens

Übungsblatt 7 zur Vorlesung
Big Data Analytics
im WiSe 2015/2016
Abgabe: 27.11.2015, 240 P

Bitte dokumentieren Sie die benötigte Bearbeitungszeit für die einzelnen Aufgaben in `bearbeitungszeit.txt`. Bitte denken Sie auch daran uns Feedback zur Veranstaltung zu geben:

`http://goo.gl/forms/B01IIDHvW2`

Bei der Abgabe von Source Code kommentieren Sie diesen bitte.
Da wir in Hive nur einen Namensraum für Tabellen nutzen, bitte fügt euren Tabellen die Gruppen Nr als Suffix hinzu.

1 Worthäufigkeiten der Wikipedia mit Hive ermitteln (180 P)

In dieser Aufgabe verwenden wir HiveQL dazu um die Worthäufigkeiten aller Wikipedia-Artikel zu bestimmen. Ziel des Aufrufs ist es eine Map mit Worthäufigkeiten für jeden Artikel zu erhalten, also in der Form:

```
12 {"by":60,"for":48,...} # für Artikel 12
13 ... # für Artikel 13
```

Dokumentieren Sie die Laufzeit des Programms. Exportieren Sie das Ergebnis als CSV in HDFS oder in das lokale Dateisystem.

1.1 Hinweise

Die Datei `wiki-clean.csv` wurde von uns bereits in HDFS importiert und befindet sich im Ordner `/user/bigdata/wiki-clean`[1]. Ein mit GZIP komprimiertes SequenceFile wurde ebenfalls angelegt und endet mit "seq.gz". Bitte achten Sie beim Anlegen des Schemas darauf, dass die Datei NICHT verschoben wird (schauen Sie sich dafür die SQL Schlüsselwörter EXTERNAL und LOCATION an).
Beim Anlegen des Schemas empfiehlt sich das Aufspalten der Zeile in Spalten mit Hilfe eines regulären Ausdrucks:

```
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe' with SERDEPROPERTIES ("input.regex" =
    ↪ "^YOUR REGEX HERE$")
```

Überprüfen Sie mit `SELECT * FROM TABLE wiki LIMIT 5;` ob Ihr Schema korrekt funktioniert. Falls Sie sich entscheiden die CSV-Datei anders zu formatieren, so löschen Sie bitte nach dem erfolgreichen durchführen Ihrer Übungen die Dateien in HDFS.
Einige Funktionen unter `https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF#LanguageManualUDF-StringFunctions` könnten für die Verarbeitung hilfreich sein.
Evtl. könnten Sie Funktionen von `https://github.com/klout/brickhouse` sinnvoll sein. Um diese nutzen zu können, verwenden Sie folgende Befehle in der Hive Shell:

```
ADD JAR /home/bigdata/brickhouse-0.7.1-SNAPSHOT.jar;
SOURCE /home/bigdata/brickhouse.hql;
```

[1] Der Header der Datei wurde ebenfalls importiert.

Abgabe:

`1-wiki-word-count.hql`	Ihre Hive-Befehle für das Anlegen der Tabelle, das Ermitteln der Worthäufigkeiten und den Export.

2 Worthäufigkeiten in Hive durch externe Skripte (90 P)

Diese Aufgabe hat das selbe Ziel wie *Aufgabe 1*. Allerdings sollen Sie nun Ihre bereits vorhandenen Python Skripte in einer TRANSFORM Clause verwenden um die Worthäufigkeiten in jedem einzelnen Artikel zu ermitteln. Sofern nötig, passen Sie hierfür Ihr Python Programm leicht an. Dokumentieren Sie die Laufzeit des Programms und vergleichen Sie diese mit dem Ergebnis von *Aufgabe 1*. Exportieren Sie das Ergebnis als CSV in HDFS oder in das lokale Dateisystem.

2.1 Hinweise

Der hierbei entstehende SQL-Aufruf sollte einfacher sein als in *Aufgabe 1*.

Abgabe:

`2-wiki-word-count.py`	Ihr Python Programm.
`2-wiki-word-count.hql`	Ihre Hive-Befehle für das Anlegen der Tabelle, das Ermitteln der Worthäufigkeiten und den Export. Der Laufzeitvergleich von *Aufgabe 1* mit dieser Aufgabe.