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

Übungsblatt 10 zur Vorlesung
Big Data Analytics
im WiSe 2015/2016
Abgabe: 31.12.2015, 420 P

Sollten Probleme auftauchen, wenden Sie sich bitte an die Mailingliste der Veranstaltung:

`bd1516@wr.informatik.uni-hamburg.de`

1 Pig Grundlagen & User Defined Functions (90 P)

In dieser Aufgabe wollen wir einige Grundlagen bei der Datenverabeitung mit Pig anhand von einfachen Beispielen erlernen:

- Schreiben Sie ein Pig-Skript, dass die Datei `/user/bigdata/numbers` einliest, den Durchschnitt aller beinhalteten Zahlen ermittelt und danach ausgibt.
- Schreiben Sie ein Pig-Skript, dass die Datei `/user/bigdata/students` einliest, pro Student die Anzahl der studierten Semester berechnet und das Ergebnis in einer neuen Datei abspeichert. Visualisieren Sie die Arbeitsweise ihres Skripts mit einem Pipe Diagram (siehe Vorlesung).
- Schreiben Sie ein Pig-Skript, dass in dem unter `/home/bigdata/1998_FIFA_World_Cup.txt` abgelegten Artikel sämtliche Vorkommnisse numerischer Zahlen bis 100 durch Text ersetzt. Implementieren Sie die Logik zur Ersetzung der Zahlen per User Defined Function in Python. Speichern Sie die neue Version des Artikels ab.

1.1 Hinweise

Sie sollten das Format des Artikels zur Fußballweltmeisterschaft im Voraus für Pig aufbereiten.
Sie finden die offizielle Dokumentation zu User Defined Functions unter `http://pig.apache.org/docs/r0.15.0/udf.html#python-udfs`.

Abgabe:

`1-compute-avg.pig`	Ihr Pig-Skript zur Berechnung des Durchschnitt.
`1-count-semester.pig`	Ihr Pig-Skript zur Berechnung der Semesteranzahl.
`1-pipe-diagram.pdf`	Ihr Pipe Diagram zum `count-semester.pig` Skript.
`1-spell-out-numbers.pig`	Ihr Pig-Skript zur Verarbeitung des Artikels.
`1-udf.py`	Ihre Python-UDF zum Ausschreiben der Zahlen.

2 NetCDF Analyse mit Pig (240 P)

In dieser Aufgabe wollen wir die Bewegungsprofile verschiedener fiktiver Menschen mit weiteren Daten korrelieren. Sie können Sich vorstellen, die Bewegungsprofile wären GPS-Daten, die mit dem Handy gesammelt wurden. Nutzen Sie Pig um das Wetter der Orte, die die jeweiligen Menschen an dem entsprechenden Tag besuchten, zu generieren und erstellen Sie damit ein Pfadprofil, welches die Wetterdaten zusätzlich enthält. Erstellen Sie dazu eine User Defined Function um von den gegeben Koordinaten auf die im NetCDF Datensatz vorhandenen Koordination zu runden.

Ausgabe soll zum einen das originale CSV erweitert um Temperatur und Niederschlag sein:

```
Name, Datum, Koordinaten, Wetter Temp, Niederschlag}
```

Zusätzlich soll ihr Pig Programm auch akkumulierte Werte pro Person in einer weiteren CSV Datei ausgeben:

```
\texttt{Name, Mittelwert Temperatur, Niederschlag}
```

Die NetCDF Daten sind bereits in HDFS importiert, Sie finde diese unter `/user/bigdata/netcdf`. Die Bewegungsdaten finden Sie unter `/user/bigdata/locations.csv`

Abgabe:

`2-weather.pig` Ihre Pig Datei die zwei CSV Dateien ausgibt.

3 Leistungsanalyse (90 P)

In dieser Aufgabe analysieren Sie die Laufzeiten von zwei von Ihnen zuvor durchgeführten Übungsaufgaben. Wählen Sie sich zwei Aufgaben nach belieben aus; wenn möglich, sollte diese mit zwei unterschiedlichen Technologien umgesetzt worden sein, bspw. Pig und Hive oder MapReduce. Sollte es im Verlauf der Aufgabe notwendig werden die Originalmessung zu wiederholen, so starten Sie diese Programme erneut. Messen Sie die Laufzeiten für Eingabedateien mit nur einem Tupel und mit dem von Ihnen zuvor genutzten Datensatz. Passen Sie die hierfür die Eingabedatei(en) unter Verwendung von `head -n 1` an und Laden Sie sie in HDFS. Mit diesen beiden Messungen können Sie den Overhead für das Starten und Beenden Ihres Programms erfassen und die Zeit pro Tupel berechnen (Workload/Zeit).

Schätzen bzw. berechnen Sie die tatsächlich genutzte Leistung für E-/A, Netzwerk und CPU. Welche Leistung hätten Sie auf unseren 5 Rechenknoten im Vergleich dazu erwartet?

3.1 Hinweise

Wir nutzen aktuell Gigabit Ethernet, eine HDD pro Knoten mit ca. 100 MiB/s und 4 Prozessoren mit 12 Kernen @2.6 GHz.

Abgabe:

`3-leistungsanalyse.pdf` Ihre Leistungsanalyse.