BDA, Praktikumsbericht 1

Gruppe: Alexander Kniesz, Maximilian Neudert, Oskar Rudolf

Repo: https://141.100.62.89:7070/#/notebook/2EA1A8WWC

Aufgabe 1

Zuerst haben wir ein gemeinsames Notebook bericht1 auf Zeppelin mit Ownern aller Gruppenmitgliedern erstellt, auf dem wir gemeinsam arbeiten können.

Wir haben als ApplicationID app-20190425171607-0336 erhalten und uns auf http://141.100.62.85:8080/ die Ressourcen angeschaut. Auffällig war, dass keine Kerne zugewiesen waren. Wenn man Testweise eine Endlosschleife mit PySpark ausgeführt hat, dann ging der Status auf Waiting. Wir sind von dem Monitor noch nicht ganz überzeugt. Der Status wirkt ziemlich träge. Aber man kann damit gut Applications abschießen die in Jobs festhängen.

Zuerst haben wir uns alle Million Song relevanten Tabellen ausgeben lassen:

```
%pyspark
spark.sql("show tables like 'msd10k*'").show(truncate=False)
```

```
+--------+------------------------+-----------+
|database|tableName               |isTemporary|
+--------+------------------------+-----------+
|default |msd10k_more_metadata    |false      |
|default |msd10k_more_metadata_row|false      |
|default |msd10k_some_metadata    |false      |
|default |msd10k_some_metadata_row|false      |
|default |msd10k_timbre           |false      |
|default |msd10k_timbre_row       |false      |
+--------+------------------------+-----------+
```

Dann haben wir die Daten gesichtet:

```
%sql
select * from msd10k_timbre limit 100
select * from msd10k_some_metadata limit 100
select * from msd10k_more_metadata limit 100
```

title	track_id	timbre_0	timbre_1	timbre_2
Doppelg��nger [Qliphothic Phantasmagoria]	TRABEFN128F92D925B	21.613	-136.784	-105.838
Doppelg��nger [Qliphothic Phantasmagoria]	TRABEFN128F92D925B	21.864	-151.802	-100.926
Doppelg��nger [Qliphothic Phantasmagoria]	TRABEFN128F92D925B	20.972	-156.087	-94.076
Doppelg��nger	TRABEFN128F92D92	22.255	-145.706	-81.169

Wir haben vorerst geprüft, ob die Timbre überall gleich lang sind

```
%pyspark
s1 = 'TRAVHPV128F933E986'
s2 = 'TRAKXYJ128F42525ED'
def get_tdur(track_id):
    not_sql_df = spark.sql("select count(timbre_0) as val from
msd10k_timbre where track_id = '{}'".format(track_id))
    s_tcount = not_sql_df.collect()[0]['val']
    not_sql_df = spark.sql("select duration as dur from
msd10k_more_metadata where track_id = '{}'".format(track_id))
    s_duration = not_sql_df.collect()[0]['dur']
    timbre_duration = s_duration / s_tcount
    return timbre_duration

d1 = get_tdur(s1)
d2 = get_tdur(s2)

print(d2 - d1)
```

Wir haben `0.0299464126059322` als Ergebnis bekommen, was bedeutet, dass die Timbre nicht gleich lang sind.