Big Data Analytics Techniques and Applications
Homework III

0556562 陳鴻君

Q1:

Program workflow:

1. import pyspark in python
2. setting config
3. load “IhaveaDream.txt”
4. word counting
5. print result

Execution commands:

```
from pyspark import SparkConf, SparkContext

conf = SparkConf().setAppName("hw3")
sc = SparkContext(conf=conf)

a = sc.textFile("IhaveaDream.txt")

c = a.flatMap(lambda line: line.split(" ")) \
        .map(lambda word: (word, 1)) \
        .reduceByKey(lambda a, b: a + b) \
        .map(lambda (a, b): (b, a)) \
        .sortByKey(False)
for x in c.collect():
        print x
```

Answers:

A.

```
(101, u'the')
(99, u'of')
(59, u'to')
(40, u'and')
(39, u'')
(36, u'a')
(32, u'be')
(27, u'will')
(24, u'that')
(23, u'is')
(21, u'in')
(20, u'we')
(20, u'as')
(19, u'freedom')
(19, u'have')
(17, u'our')
(17, u'from')
(15, u'I')
(13, u'Negro')
(13, u’not')
```

B.
freedom
Negro
dream

These three words are the main words in this article and appear most except meaningless words.

Q2:

Program workflow:

1. import pyspark and pysparkSQL
2. setting context
3. read csv
4. select ‘passenger_count’, ‘payment_type’ and filter passenger_count > 0
5. groupBy ‘payment_type’ and calculate mean
6. show result

Execution commands:

```
from pyspark.sql import SQLContext
from pyspark import SparkContext, SparkConf

conf = SparkConf().setAppName("hw3_Q2")
sc = SparkContext(conf = conf)
sqc = SQLContext(sc)

a = sqc.read.format("com.databricks.spark.csv") \
        .options(header = 'true', inferschema = 'true') \
        .load("yellow_tripdata_2016-08.csv")

b = a.select('passenger_count', 'payment_type').filter(a.passenger_count
> 0)
c = b.groupBy('payment_type').mean()
c.show()
```

Answers:

```
+------------+--------------------+-----------------+
|payment_type|avg(passenger_count)|avg(payment_type)|
+------------+--------------------+-----------------+
|           1|  1.6403510531217604|              1.0|
|           2|  1.7164726170150806|              2.0|
|           3|  1.2989506430793518|              3.0|
|           4|  1.3307480786857506|              4.0|
|           5|                 1.0|              5.0|
+------------+--------------------+-----------------+
```

Q3:

Program workflow:

1. run on yarn platform
2. run on local platform

Execution commands:

```
time spark-submit --master yarn hw3.py > Q1.txt
time spark-submit --master local[*] hw3.py > Q1.txt
```

Answers:

yarn ->

```
real0m17.042s
user0m28.955s
sys 0m2.390s
```

local ->

```
real0m6.622s
user0m12.682s
sys 0m1.704s
```