Big Data Analytics Hw3

0556555 林哲宇

Q1.

Workflow:

1.Load the file ‘IhaveaDream.txt’ on hdfs
2.flat the content with delimiter “ ”
3.map every word into a value key
4.reduce by value key
5.do the sorting (make me easier to find key words)

Execution commands:

spark-submit hw3.py (other config are set in Q3)

Answer:

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

The Keyword of the topic should be “The Nation Negro’s Freedom”.

Q2.

Workflow:

1. read csv file
2. select needed columns and filter it to get rid of null value
3. group it up by payment type
4. calculate the mean value of each group

Execution commands:

spark-submit --packages com.databricks:spark-csv_2.10:1.5.0 hw3_2.py

Answer:

Payment Type	Avg(Passenger_Count)
1	1.656962440784346
2	1.7037601746358282
3	1.263854792373988
4	1.316774819124338
5	1.0

Payment type 2 have the most passenger.

Q3.

Workflow:

same as Q1, but added time function to measure the execution time

Execution commands:

local: spark-submit —master local[*] hw3.py
yarn: spark-submit —master yarn hw3.py

Answer:

The execution time of yarn cluster is faster than local-worker, but total time is slower.

Method	Execution time	time-real	time-sys
local all	0.04	8.296	3.245
yarn cluster	0.03	15.777	1.030

Discussion

1. Though finding the word-frequency of an article, you still can't ensure that all the words' importance. For Q1, 'the','is','am','are' are highly-occurred. But in real world, those words are not important.
2. I use **filter** to get rid of data loss in Q2.
3. spark-submit is useful, but seeking for dependency problem is harsh.
4. For Q3, though yarn cluster's execution time is faster than local worker. But I noticed that the task completion of local worker is faster, but the task manager is slower. Yarn cluster may have better task manage ability, but task completion is dropped by I/O interrupt.