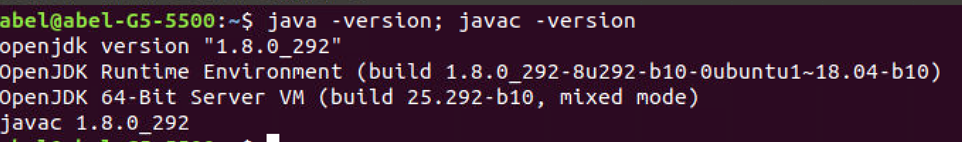
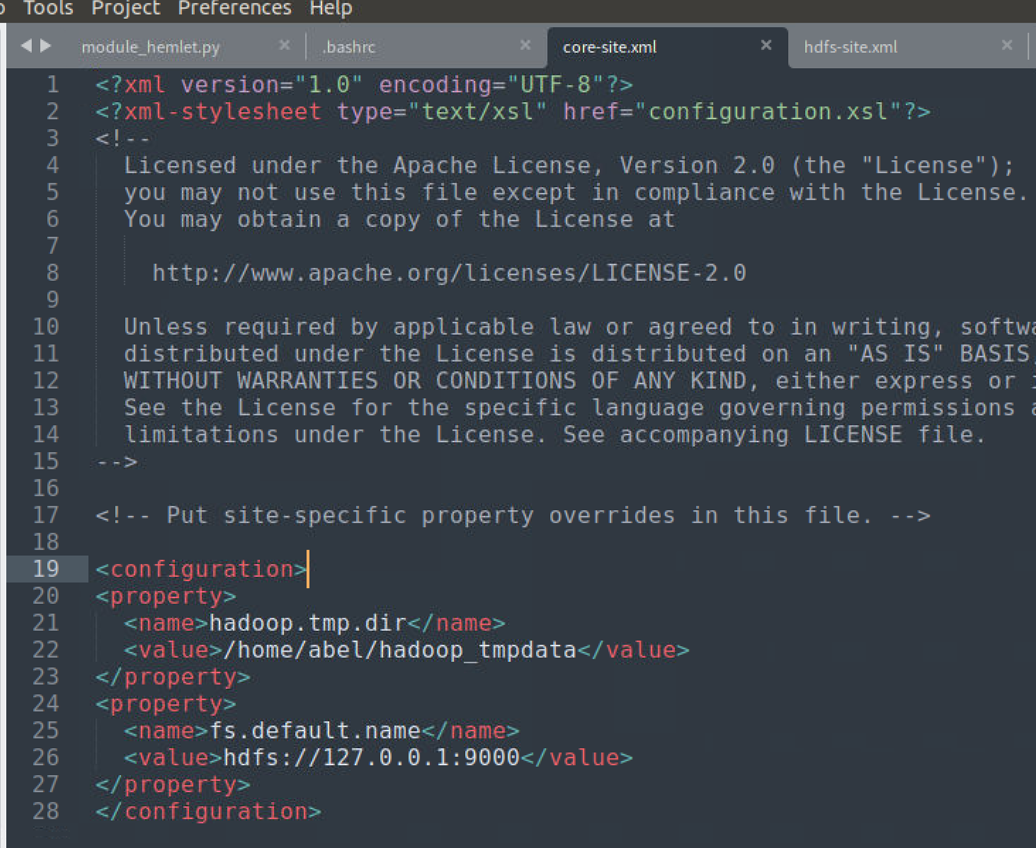
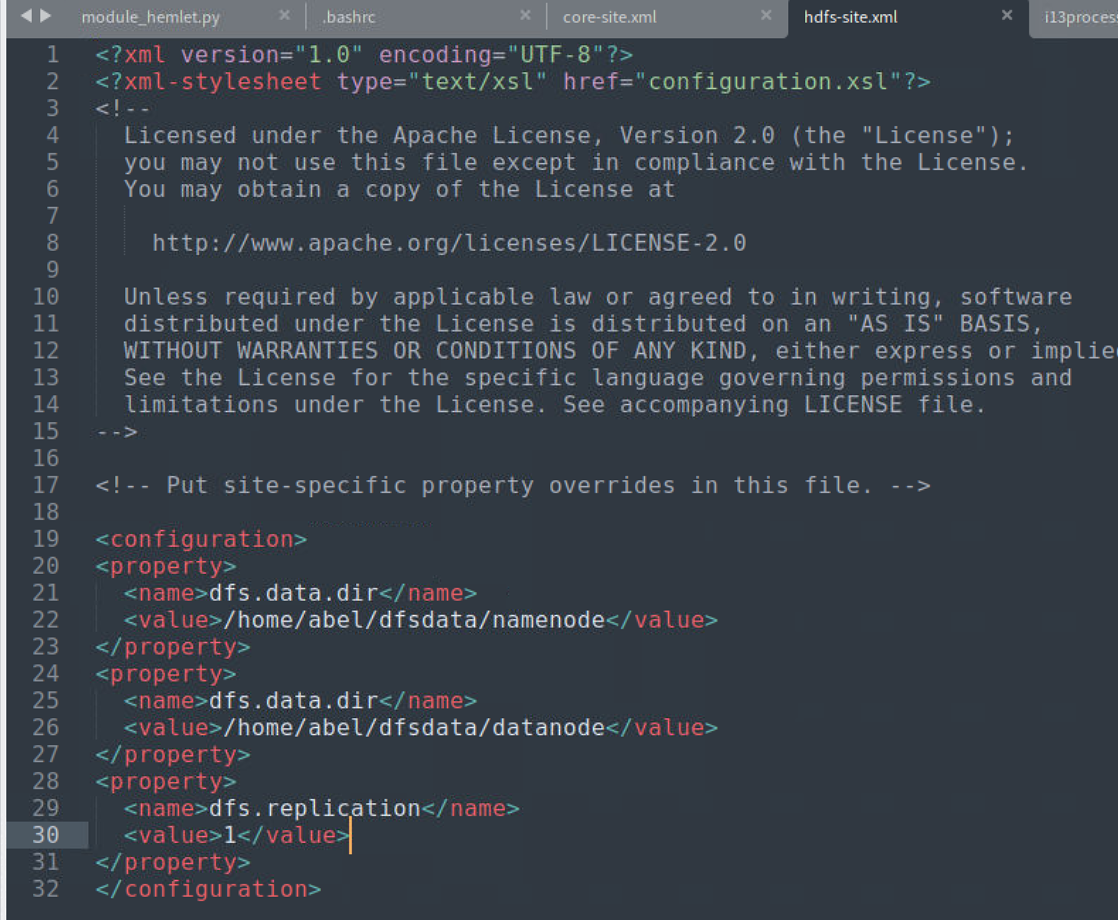
Small Project

Q1.

a.

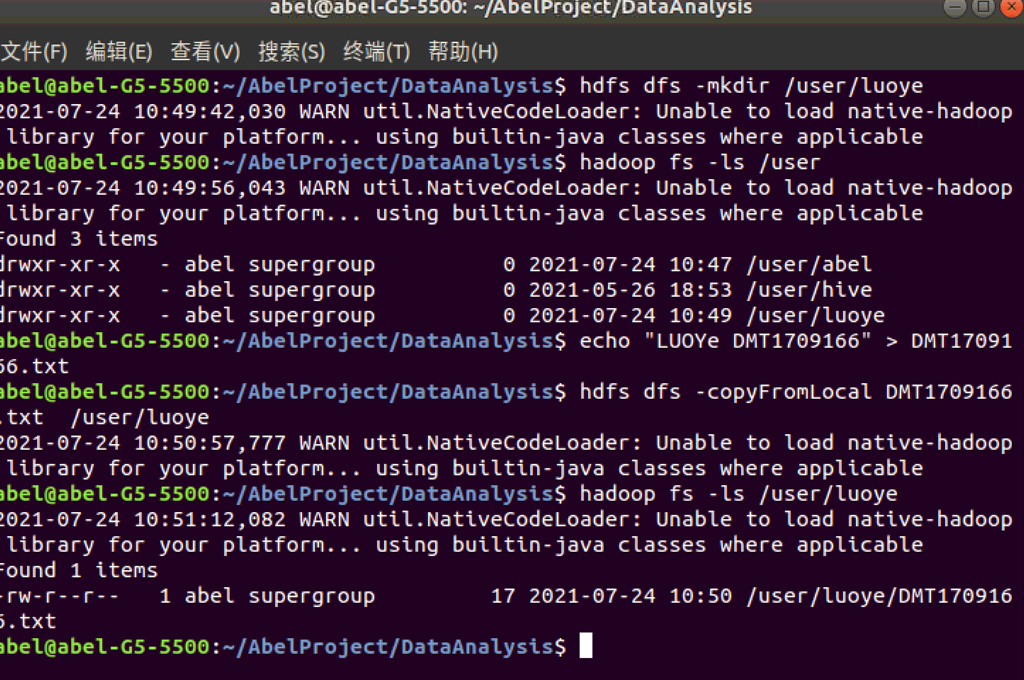
java environment used by hadoop:configuration files that you’ve modified to setup the pseudo-distributed mode :

b.

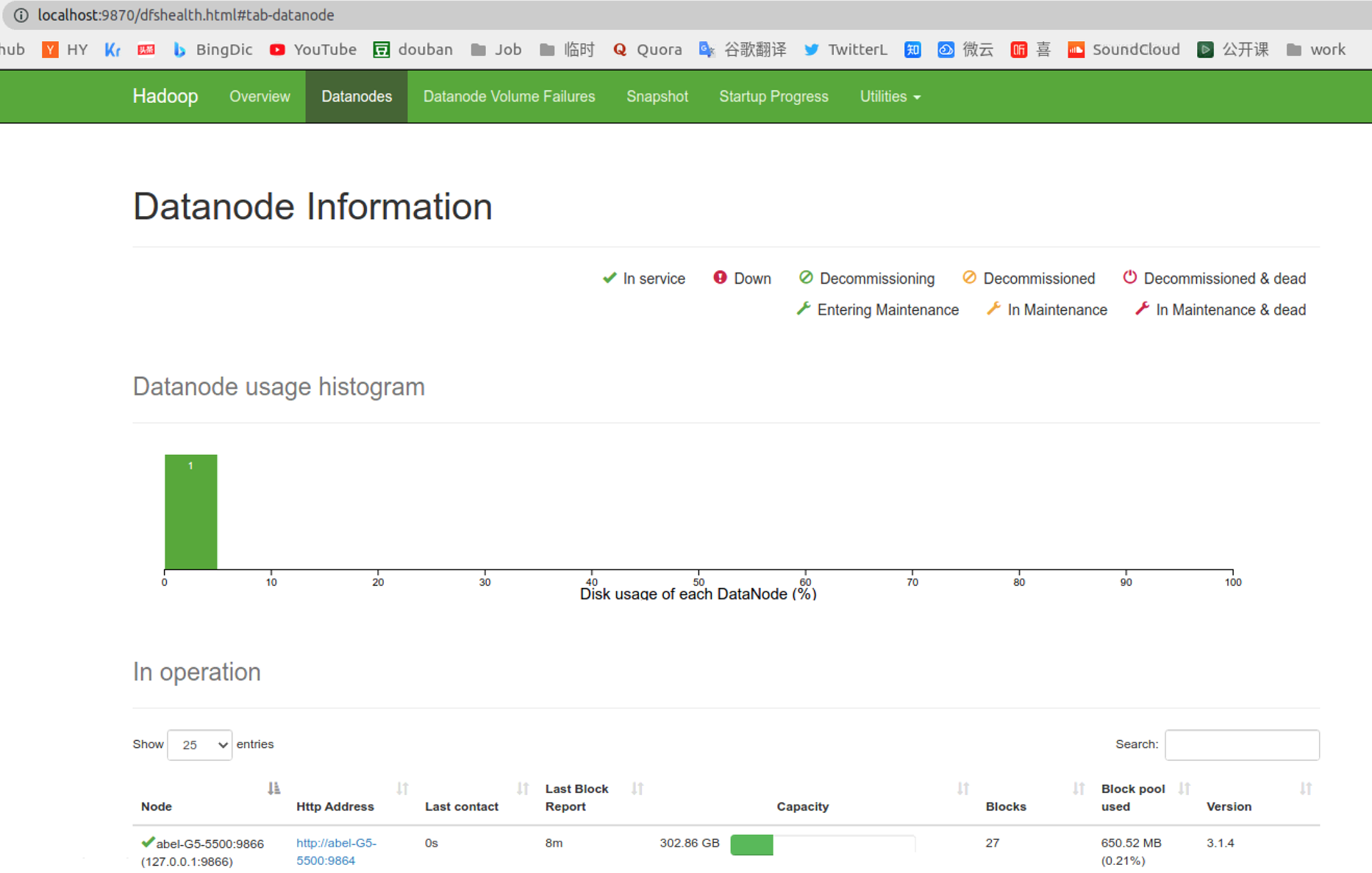
echo "LUOYe DMT1709166" > DMT1709166.txt

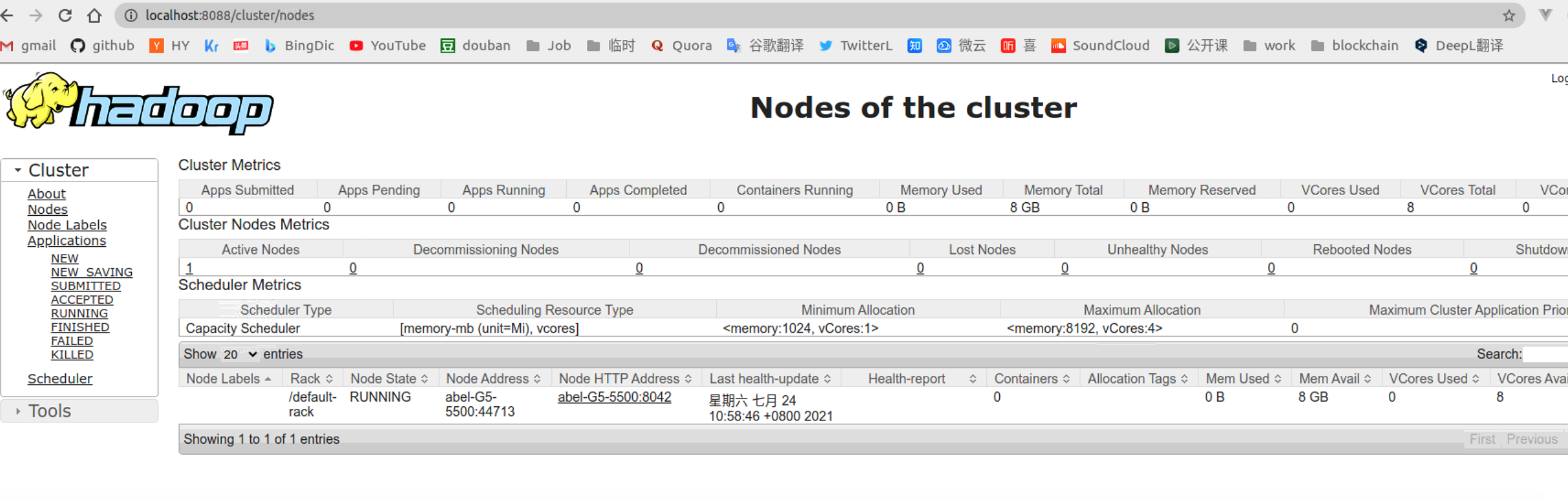
hdfs dfs -mkdir /user/luoye

hdfs dfs -copyFromLocal DMT1709166.txt /user/luoye



c.





**Q2**

2.1.a

map function of (a) task:

mapping:

(itemx in x) (itemy in y) -> (itemx, itemy) in (x,y)

reduce:

Many itemx+itemy -> min number of the sum(itemx+itemy)

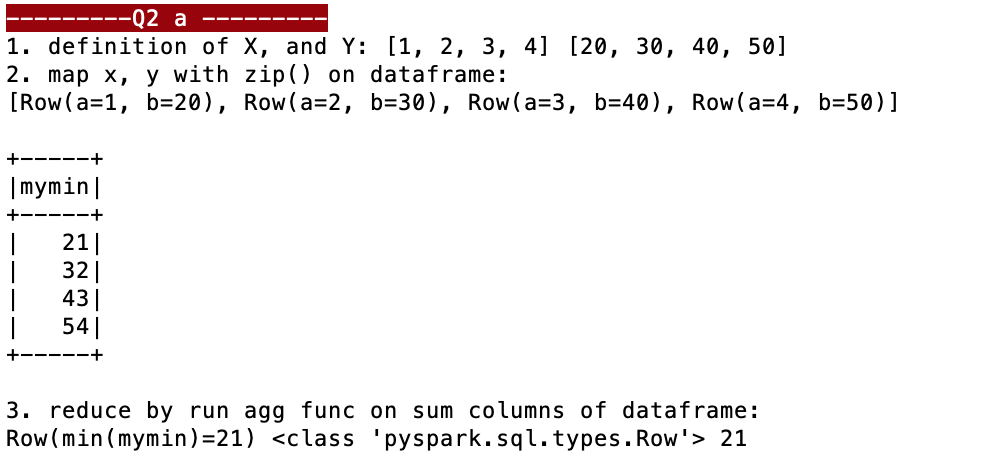
code as following:

2.1.b

2.2.a

map function of (a) task:

mapping:

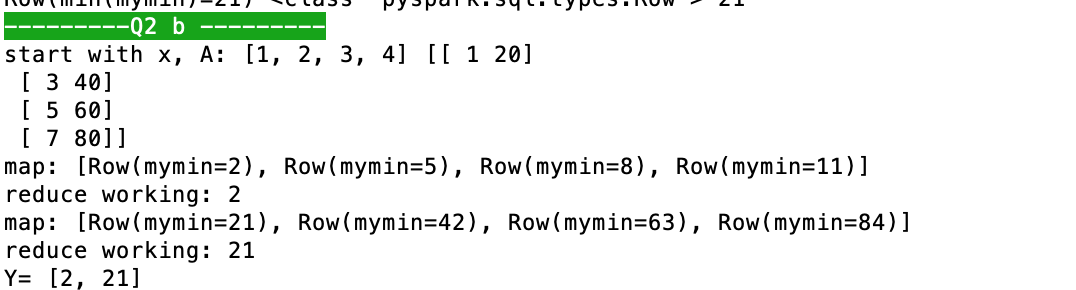
(itemx in x) (columns\_of\_y in y) -> (itemx, columns\_of\_y) in (x,y)

reduce:

Many itemx+columns\_of\_y -> min number of the sum(itemx+columns\_of\_y)

code as following:

2.2.b



**Q3**

from pyspark import SparkContext, SparkConf

from pyspark.sql import HiveContext, Row

from pyspark.sql.types import IntegerType

import json

import sys

inputFile = 'flights2008.csv'

inputFile = '/Users/abel/Downloads/spare\_time/working/b13590\_1.8k\_spark\_7月28日/flights2008.csv'

conf = SparkConf().setAppName("SparkSQLAirTransit")

SparkConf().set("spark.sql.legacy.timeParserPolicy","LEGACY")

sc = SparkContext.getOrCreate(conf=conf)

hiveCtx = HiveContext(sc)

print("Loading flights from " + inputFile)

input = hiveCtx.read.option("header",True).csv(inputFile,inferSchema =True)

input.registerTempTable("air\_transit")

myair\_transits = hiveCtx.sql("SELECT \* FROM air\_transit LIMIT 3")

print('myair\_transits:' )

for item in myair\_transits.collect():

print(item, '\n')

print('1. column=',len(myair\_transits.columns))

myair\_transits = hiveCtx.sql("SELECT count(\*) FROM air\_transit ")

print('1. rows : ', myair\_transits.collect())

print('2. print out the schema of this dataframe')

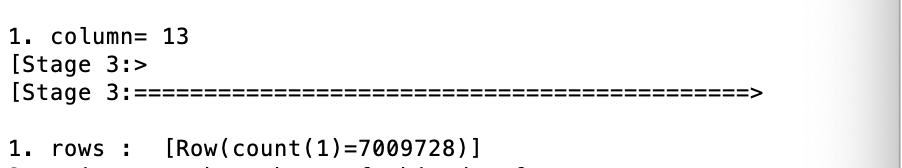
input.printSchema()

print('3. List the distinct carriers of in the dataframe:')

myair\_transits = hiveCtx.sql("SELECT distinct UniqueCarrier FROM air\_transit ")

print(myair\_transits.collect())

print('4. How many flights in total are there in January:')

myair\_transits = hiveCtx.sql("SELECT count(\*) FROM air\_transit where Month=1 ")

print(myair\_transits.collect())

print('5. the total number of flights by each of the carriers')

myair\_transits = hiveCtx.sql("SELECT UniqueCarrier, count(\*) FROM air\_transit group by UniqueCarrier ")

print('UniqueCarrier', ' count(\*)')

for row in myair\_transits.collect():

print(row[0], row[1])

print('6. Count the total number of flights in the first half year (month 1-6) by each of the carriers.')

myair\_transits = hiveCtx.sql("SELECT UniqueCarrier, count(\*) FROM air\_transit where Month <= 6 group by UniqueCarrier ")

print('UniqueCarrier', ' count(\*)')

for row in myair\_transits.collect():

print(row[0], row[1])

