gedit ubermapper.py

#!/usr/bin/python

import sys

def mapper():

for line in sys.stdin:

dispatching\_base\_number, date, active\_vehicles, trips = line.strip().split('\t') print(f"{0}\t{1}\t{2}".format(dispatching\_base\_number,date,trips)

mapper()

gedit uber.txt

B001,2022-01-01,10,50

B001,2022-01-02,12,55

B002,2022-01-01,8,40

B002,2022-01-02,10,45

B001,2022-01-03,14,60 UBER

B002,2022-01-03,12,50

cat uber.txt | python ubermapper.py

gedit uberreducer.py

#!/usr/bin/python

import sys

def reducer():

current\_key = None

current\_max\_trips = 0

for line in sys.stdin:

dispatching\_base\_number, date, trips = line.strip().split('\t')

if dispatching\_base\_number=='dispatching\_base\_number':continue

if current\_key is not None and current\_key != dispatching\_base\_number: print"{0}\t{1}\t{2}".format(current\_key,date,current\_max\_trips)

current\_max\_trips = 0

current\_max\_trips = max(current\_max\_trips, int(trips))

current\_key = dispatching\_base\_number

if current\_key is not None:

print"{0}\t{1}\t{2}".format(current\_key,date,current\_max\_trips)

reducer()

cat uber.txt | python ubermapper.py | sort | python uberreducer.py

hdfs dfs -put uber.txt uber.txt