DS-200$^{Q\&As}$

Data Science Essentials

Pass Cloudera DS-200 Exam with 100% Guarantee

QUESTION 1

You have a data file that contains two trillion records, one record per line (comma separated). Each record lists two friends and unique message sent between them. Their names will not have commas.

Michael, John, Pabst, Blue Ribbon Tiffany, James, BMX Racing John, Michael, Natural Lemon Flavor

Analyze the pseudo code examples below and determine which set of mappers and reducers in the below pseudo code snippets will solve for the mean number of messages each user sends to all of the friends?

For example pseudo code may have three friends to whom he sends 6, 10, and 200 messages, respectively, so Michael\\'s mean would be (6+10+200)/3. The solution may require a pipeline of two MapReduce jobs.

A. def mapper1 (line): key1, key2, message = line.split (` , \\') emit ((key1, key2) , 1) def reducer1(key, values): emit (key, sum(values)) def mapper2(key, value): key1, key2 = key / / unpack both friends name into separate keys emit (key1, value)

def reducer2(key, values):

emit (key, mean (values))

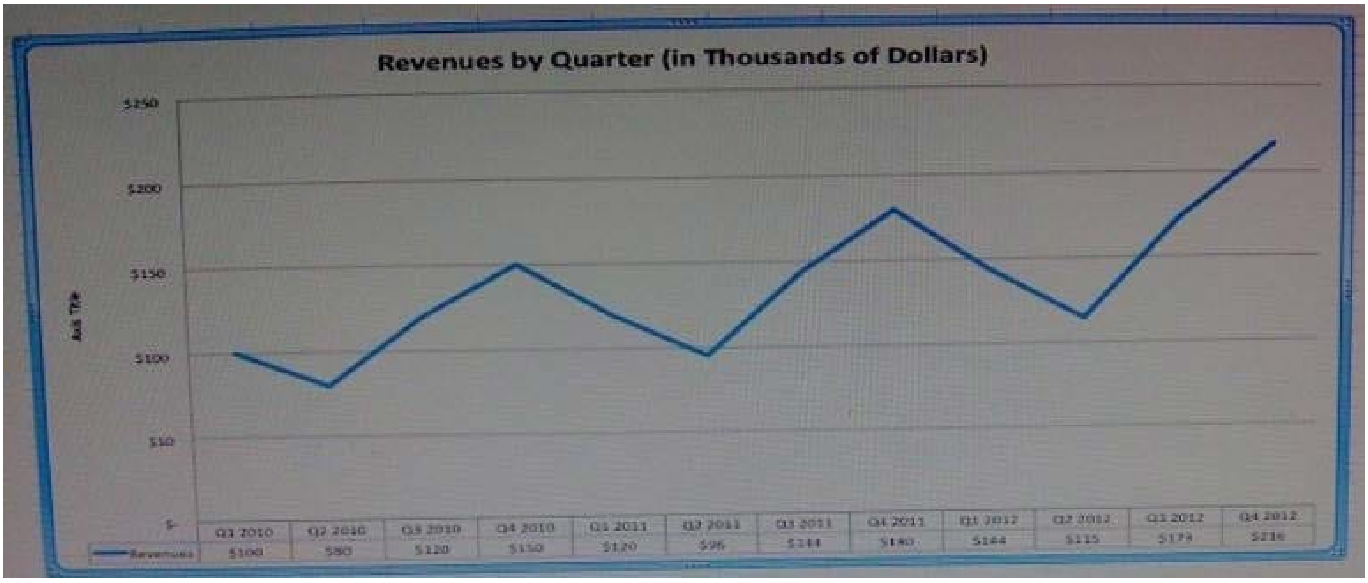
B. def mapper1 (line): key1, key2, message = line.split (` , \\') emit ((key1, key2) , 1) emit ((key1, key2) , 1) def reducer1(key, values): emit (key, sum(values)) def mapper2(key, value): key1, key2 = key / / unpack both friends name into separate keys emit (key1, value) def reducer2(key, values): emit (key, mean (values))

C. def mapper1 (line): key1, key2, message = line.split (` , \\') emit ((key1, key2) , 1) emit ((key1, key2) , 1) def reducer1(key, values): emit (key, sum(values))

D. def mapper (line) : Key1, key2, message = line.split (` , \\') Sort (key1, key2) / / a fiven pair will always be sorted the same Emit ((key 1, key2), 1) Def reducer1(key, values) : Emit (key, sum (values)) Def Mapper2 (key, value) Key1, key2 = key / / unpack both friends names into separate keys Emit (key1, value) Emit (key2, value) Def reducer2(key, values); Emit (key, mean (values))

Correct Answer: B

QUESTION 2



Assuming the trends shown in this chart continue, what would we expect the value of the revenue to be in Q1 of 2013?

A. $125,000

B. $170,000

C. $220,000

D. $250,000

Correct Answer: A

QUESTION 3

You have a large file of N records (one per line), and want to randomly sample 10% them. You have two

functions that are perfect random number generators (through they are a bit slow):

Random_uniform () generates a uniformly distributed number in the interval [0, 1] random_permotation (M)

generates a random permutation of the number O through M -1.

Below are three different functions that implement the sampling.

Method A

For line in file: If random_uniform ()

Method B

i = 0

for line in file:

```
if i % 10 = = 0;
print line
i += 1
```

Method C

```
idxs = random_permotation (N) [: (N/10)]
i = 0
for line in file:
if i in idxs:
print line
i +=1
```

Which method is least likely to give you exactly 10% of your data?

A. Method A

B. Method B

C. Method C

Correct Answer: B

QUESTION 4

Which best describes the primary function of Flume?

A. Flume is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with an infrastructure consisting of sources and sinks for importing and evaluating large data sets

B. Flume acts as a Hadoop filesystem for log files

C. Flume Imports data from SQL/relational database into your Hadoop cluster

D. Flume provides a query languages for Hadoop similar to SQL

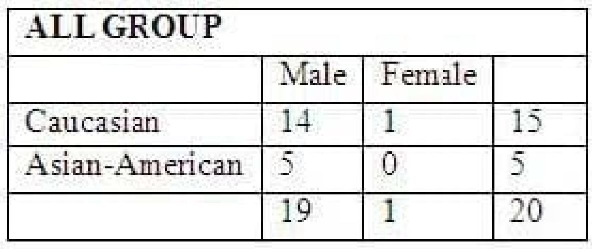
E. Flume is a distributed server for collecting and moving large amount of data into HDFS as it\\'s produced from streaming data flows

Correct Answer: D

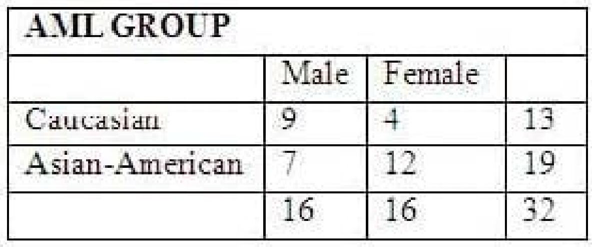
QUESTION 5

There are 20 patients with acute lymphoblastic leukemia (ALL) and 32 patients with acute myeloid leukemia (AML), both variants of a blood cancer.

The makeup of the groups as follows:



ALL GROUP			
	Male	Female	
Caucasian	14	1	15
Asian-American	5	0	5
	19	1	20



AML GROUP			
	Male	Female	
Caucasian	9	4	13
Asian-American	7	12	19
	16	16	32

Each individual has an expression value for each of 10000 different genes. The expression value for each gene is a continuous value between -1 and 1.

With which type of plot can you encode the most amount of the data visually?

You choose to perform agglomerative hierarchical clustering on the 10,000 features. How much RAM do you need to hold the distance Matrix, assuming each distance value is 64-bit double?

A. ~ 800 MB

B. ~ 400 MB

C. ~ 160 KB

D. ~ 4 MB

Correct Answer: B

DS-200 VCE Dumps DS-200 Practice Test DS-200 Braindumps