

CSE 3244: Data Management in the Cloud
Exam #1: Autumn 2020

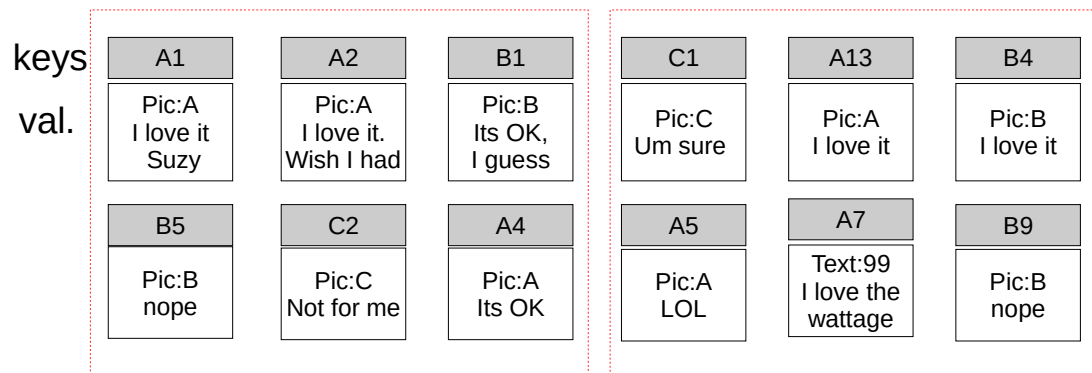
Score: _____ of 150

Problem 1 [40 points total; 4 points each].

1. A petabyte is 6 times larger than a gigabyte. ____True ____ False
2. Structured, relational data can comprise a part of unstructured data . ____True ____ False
3. Virtualization allows multiple operating systems to run atop the same hardware. ____True ____ False
4. Data scientists write map & reduce code that manages (1) where tasks run and (2) how data is partitioned.
____True ____ False
5. For term co-occurrence, the all-pairs algorithm discussed in class is a _____ implementation.
____Holistic ____ Algebraic ____ Distributive
6. For term co-occurrence, the all-pairs algorithm outputs smaller objects than the stripes algorithms.
____True ____ False
7. The key-value pairs emitted by a map task are sent directly to reduce tasks. ____True ____ False
8. Map tasks run concurrently (at the same time) as reduce tasks to ensure high throughput.
____True ____ False
9. A network switch connects disks to memory within a blade ____ True ____ False
10. The Y-Axis on a Throughput-Capacity curve captures throughput for map tasks. The X-axis distinguishes map-reduce configurations. ____ True ____ False

Problem 2: [66 points; 11 points each]

The following questions reference the map reduce job in the figure below. It aims to capture “love” sentiment expressed in comments about pictures posted online. Dotted lines indicate which key value pairs map to each map task. The Map Task boxes execute the code within them. There are two map tasks, each has an output.



Node #1

Node #2

Map Function

Map Function

Mapper
Output

```
Map class (k,v) {
    name = V.getFirstLine()
    If (v.contains("love")) {
        Emit (name,1)
    }
    Else {
        Emit (name,0)
    }
}
```

```
Map class (k,v) {
    name = V.getFirstLine()
    If (v.contains("love")) {
        Emit (name,1)
    }
}
```

Shuffle phase



Reducer
Input

Please write your answers here.

1. Look closely at this example, it does not reflect a valid execution of the map stage in map reduce. Why?

2. What are the outputs of Map Instance that processes key A1? (check all that apply)

3. Consider the Map Instance that processes key A1, the inputs to a combiner executing on the same node would be? (check all that apply)

4. Fill in the blanks below for the reduce stage. The goal is to compute percentage of each pic's comments that express "love" sentiment (i.e., comments including "love" over total comments).

```
Reduce Class (k, [v]) {  
    int sum = 0;  
    int count = 0;  
    for each v in V  
        _____  
        _____  
    Emit (k, _____)  
}
```

5. True or false: Given the map and reduce functions above. If the number of key value pairs increased 40X, we would need to rewrite the code to use more mappers/reducers?

___ True ___ False

6. True or false: Given the map and reduce functions above. If we want to compute the share of "love" attributed to each pic (i.e., loved for picA over total loved for all pics), we would need to rewrite the map and/or reduce function. ___ True ___ False

Problem 3: [14 points]

- _____ True _____ False The number of map tasks is determined by the number of blades
- _____ True _____ False The number of map instances is determined by the number of blades
- _____ True _____ False The number of NameNodes in HDFS is determined by the number of key-value inputs
- _____ True _____ False The number of reduce tasks is determined by the number of unique intermediate keys
- _____ True _____ False The number of reduce tasks is determined by the number of map tasks

Contextual information:

- In your company's datacenter, blades have 500 GB disk space for storage and 24 GB RAM
- Blades in a rack are connected to a 20 port CISCO switch
- Each disk provides 500 MB/s peak throughput, DDR3 DRAM provides 15 GB/s peak throughput, the switch transmits at 1 GB/s. You have access to only 1 rack.
- Alternatively, a cloud data center provides 50 GB free disk storage and 12 GB RAM and a two-tier hierarchy for networking speeds. Rack switches provide 10 GB/s over 100 ports. Racks are connected with 500 MB/s over 100 ports. SSD speeds are 2 GB/s. DDR4 DRAM offers 30 GB/s.

Problem 5 [30 points].

1. Which configuration provides the maximum throughput for map tasks?
2. If map tasks access input data stored in-memory on machines attached the same network switch, which configuration provides highest throughput? _____ Your company _____ Cloud data center
3. Which configuration offers the greatest storage capacity for a single map task? _____ Your Company _____ Cloud data center
4. You want to save money by buying cloud blades that use slower DDR3 RAM (like your your local company). Assuming map-tasks require more than 15,000 MB storage, will this affect throughput?
_____ Yes _____ No
5. How much would upgrading your company's rack-level switch improve throughput?
_____ >2X _____ 1.5X _____ 1X (flat) _____ 0.5X (reduce throughput)