

EE4221 Lecture 7 Questions

Analytics for Unstructured Data - MapReduce and Hadoop

Q1. Consider two tasks:

Task A: `foo(a)`

Task B: `bar(b)`

If `x:=foo(a)+bar(b)`, and `a:=b+c+d`, can Tasks A and B be run in parallel by using MapReduce?

- a. Yes
- b. No **
- c. Not sure

Q2. Consider these two threads:

Thread 1:

```
void foo() {  
    eax = mem[x];  
    inc eax;  
    mem[x] = eax;  
    ebx = mem[x];  
    mem[y] = ebx;  
}
```

Thread 2:

```
void bar() {  
    eax = mem[y];  
    inc eax;  
    mem[y] = eax;  
    eax = mem[x];  
    add eax, 3;  
    mem[x] = eax;  
}
```

If these two threads are run in two different CPUs, can they be run in parallel?

- a. Yes
- b. No **
- c. Not sure

Q3. In an application, we want to find how many French words are used in an extremely large (but static) English dataset. Can we use MapReduce in this application?

- a. Yes **
- b. No
- c. Not sure

Q4. In an application, we want to find how many French words are used in an extremely large English dataset. The dataset is changing from time to time. Can we use MapReduce

in this application? {Hint: go to web and download this paper: <http://storageconference.org/2010/Papers/MSST/Shvachko.pdf>. Read section III of the paper.}

- a. Yes
- b. No **
- c. Not sure

Q5. In an application, we want to find how many French words are used in an extremely large English dataset. New contents are continuously appended to the dataset. Can we use MapReduce in this application?

- a. Yes **
- b. No
- c. Not sure

Q6. Is Amazon supporting Hadoop? If yes, what is the name of this service?

Amazon Elastic MapReduce (Amazon EMR)