

ID2221 - Data Intensive Computing

Oct. 18, 2019 - 8:00-12:00

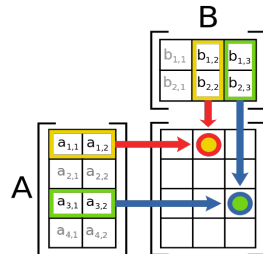
Examiner: Amir Payberah (phone num: 072-55 44 011)

No aids allowed

A: $x \geq 45$, **B:** $40 \leq x < 45$, **C:** $35 \leq x < 40$, **D:** $30 \leq x < 35$, **E:** $25 \leq x < 30$, **F:** $x < 25$

1. Distributed file systems such as GFS store files in chunks. The size of a chunk has implications on the system's performance. What are the advantages and disadvantages of storing files in large chunks? (5 points)
-
2. Explain briefly the similarities and differences of document-based and column oriented databases. What is the difference between storing data in row model and column model? When is better to use the column model? (5 points)
-
3. Assume we want to implement a MapReduce code to multiply a one-dimensional matrix with a two-dimensional matrix. Briefly explain what map and reduce functions should do. (5 points)

Reminder. The matrix product of matrices **A** and **B** is a third matrix **C**, where $\mathbf{C} = \mathbf{AB}$. If **A** is of shape $m \times n$ and **B** is of shape $n \times p$, then **C** is of shape $m \times p$, such that:

$$c_{ij} = \sum_k a_{ik} b_{kj}$$


-
4. Draw the lineage graph for the following code and mention which connections are narrow and which ones are wide? (5 points)

```
val a = sc.parallelize(...)
val c = sc.parallelize(...)
val e = sc.parallelize(...)
val b = a.groupby(...)
val d = c.map(...)
val f = d.union(e)
val g = b.join(f)
```

-
5. Explain how does using DataFrame in Spark improve the programming performance compared to RDD. What is the relation between DataFrame and DataSet in Spark? (5 points)

-
6. Explain briefly how does spark join two tables, if (5 points)

- (a) both tables are so big that none of them can be loaded into memory of one computer
- (b) one table is big and the other one is small, such that only the small one can be loaded in memory of one computer.

-
7. How can a batch system be used to process streams, and how can a streaming system be used to process batches? (5 points)

-
8. Explain how Kafka provides scalability and fault tolerance? (4 points)

-
9. Compare the graph processing models in Pregel, GraphLab, X-Stream. (6 points)

-
10. Assume we have two types of resources in the system, i.e., CPU and Memory. In total we have 10 CPU and 20GB RAM. There are two users in the systems. User 1 needs $\langle 1CPU, 4GB \rangle$ per task, and user 2 needs $\langle 2CPU, 1GB \rangle$ per task. How do you share the resources fairly among these two users, considering the asset fairness and DRF. (5 points)