# CS 544 (Fall 2025): In-Person Quiz 2 (Version A)

Full Name: ____________________________________________________

Student ID Number: ______________________________________________

1. Which SQL clause is used to filter rows before a GROUP BY is applied?
   **(A) FILTER (B) HAVING (C) BEFORE (D) PROJECT (E) WHERE**
2. Which type of database usually has a column-oriented layout?
   **(A) OLAP (B) OLTP**
3. If you want FEWER logical blocks in a NameNode's block map, what should you do?
   **(A) increase block size (B) decrease block size (C) decrease replication factor**
4. If you're trying to implement the equivalent of a GROUP BY in a MapReduce job, where should you specify the column by which to group?
   **(A) keys emitted by map (B) values emitted by map**
   **(C) keys emitted by reduce (D) values emitted by reduce**
5. Which Spark operation is a transformation?
   **(A) take (B) toPandas (C) collect (D) map**
6. Your Spark cluster has 10 machines, each with 8 GB of RAM and 4 CPU cores. Your RDD (when materialized) consumes 100 GB of RAM and has 100 partitions. If we call .map(...) on the RDD, how many tasks can we run concurrently?
   **(A) 10 (B) 40 (C) 80 (D) 100**
7. Which one is a pseudo file system?
   **(A) procfs (B) ext4 (C) NFS (D) tmpfs**
8. Which of the following flags should you use while running a docker container if you want files that you created or modified inside the docker container to be accessible even after terminating the container?
   **(A) -d (B) -p (C) -q (D) -v (E) -l**
9. What do you call a section of code where certain interleavings with other code would be a problem?
   **(A) lock (B) thread (C) race condition (D) critical section (E) deadlock**
10. Suppose you have launched a docker container running with port forwarding option “127.0.0.1:8870:9870” and Hadoop installed. Then, you create an SSH tunnel using -L localhost:7870:localhost:8870. If you want to send requests to WebHDFS from your laptop, what port should you use?
    **(A) 5000 (B) 7870 (C) 8870 (D) 9000 (E) 9870**

**Question 11 (5 points).** Draw an HDFS cluster with a NameNode (NN), 4 DataNodes (DN1 to DN4), and one client, each represented as a labeled rectangle. The client is writing "abcd" to a new, single-block file with 3x replication. Draw arrows between nodes in the cluster, and label them as "M" (metadata transfer) or "D" (data transfer). Illustrate the physical blocks where they will reside, as quoted strings. Write "BM" on any node(s) with a full-copy of the in-memory block map.

**Question 12 (5 points).** Consider the following Spark code:

```
A = ... # RDD of all ints from 0 to 1 million, in 10 partitions
B = A.map(lambda x: 1/(1+x))
C = A.mean()
D = A.filter(lambda x: x % 5 == 0)
E = D.take(5)
```

Finish drawing a directed graph of the above 5 things (A to E). Represent RDDs as boxes and materialized values (ints, Python lists, etc.) as circles. For RDDs, write the number of partitions beneath the box. Label edges to indicate transformation (T) or action (A).