MapReduce Recap Quiz:

1. MapReduce is a programming model/framework, while Hadoop is an implementation of MapReduce.

2. What do we mean with shuffling in a MapReduce context?
    a. Transferring data through the network.
    b. Splitting data into partitions.
    c. Rearranging data before sorting it.

3. Which functions does a user need to define in MapReduce?
    a. reduce()
    b. pre-aggregate()
    c. shuffle()
    d. map()
    e. sort()

4. An embarrassingly parallel program/algorithm..
    a. ..splits processing into independent tasks.
    b. ..requires a global variable (state) among the parallel processes.
    c. ..requires communication among the parallel processes.

5. Match the answers to the following questions:
    a. What is the size of an HDFS block by default? 64MB
    b. Why are HDFS blocks replicated 3 times? For fault-tolerance and load balancing reasons.
    c. What is the characteristic of HDFS wrt updates? The size of an HDFS block is equal to the corresponding disk blocks size.

Spark Recap Quiz:

1. A dataflow is a directed acyclic graph (DAG) where nodes represent operations and edges represent the data flowing from one operator to the other.
    a. False
    b. True

2. A dataflow engine receives as input a [logical] plan and creates a [physical] plan. It includes both [logical] optimizations, such as [operator re-ordering], and [physical] optimizations such as [join algorithm selection]. It runs on a [cluster] in [parallel] while handling [failures].

3. The following code in Spark will return a list of (10, 20, 30, 40).

    ```
    data = ArrayList(1, 2, 3, 4)
    values = sc.parallellize(data).map(number -> number*10)
    ```
    a. True
    b. False

4. Spark is faster than Hadoop because:
    a. It provides a functional programming interface.
    b. It uses modern hardware and accelarators.
    c. It can re-use data computed by certain operations.
    d. It uses hash partitioning to shuffle the data.
    e. It keeps data in memory (if possible) instead of saving intermediate results on disk.

5. Match the following:
    a. A Spark job is → a set of tasks executed as a result of an action operation.
    b. A stage in Spark is → a set of tasks in a job executed in parallel without any network communication.
    c. A task in Spark is → a unit of work over one partition sent to one executor.

Cross-platform data processing:

1. In cross-platform data processing we may use multiple systems for processing a single query for different reasons. How do we call each of the following cross-platform cases?
    a. The data is stored in different systems → Polystore cross-platform processing
    b. We want to optimize performance → Opportunistic cross-platform processing
    c. The system where the data is stored does not have a desired functionality → Mandatory cross-platform processing

2. What are the problems with a cost-based optimizer in a cross-platform setting?
    a. The cost functions assume linear behaviour.
    b. The cross-platform system may not have access to data statistics.
    c. It is very hard to define the different cost functions.
    d. Data movement is costly.
    e. It is hard to fine-tune the coefficients of the cost functions.
    f. Query optimization is very time consuming.

3. Wayang allows you to write a pipeline once and run it on multiple execution engines without rewriting it.
    a. False
    b. True

4. In a cross-platform setting, what does “data movement” refer to?
    a. Copying files from one location to another
    b. Moving data between different execution platforms during a workflow
    c. Changing the type of the data

5. Which scenario would likely trigger data movement in Wayang?
    a. Printing a small number of records to the console
    b. Reading records from a table in Postgres, filtering the records, transforming them into numerical features, and training an ML model in Tensorflow
    c. Reading a text file and immediately writing it to the same folder
    d. Joining a PostgreSQL table with a CSV file on HDFS, then executing on Spark

ML lifecycle recap

1. When splitting a dataset into train and test sets, it's important to:
    a. a test dataset is not always necessary
    b. shuffle the dataset and pick samples at random
    c. always pick 80% of the data for training and 20% for testing
    d. depending on the data: either shuffle datapoints or maintain order for timeseries data with temporal dependencies
2. Data leakage can be avoided by:
    a. scaling the dataset before splitting into train and test sets
    b. cleaning the data
    c. chaining transformations as part of a pipeline
    d. outlier detection
3. Which of the following is true about sklearn pipelines?
    a. Transformers are a type of Estimators.
    b. Pipelines ensure that the same data transformations are applied consistently during both training and prediction, reducing the risk of data leakage
    c. You need to call fit and predict for every step in the pipeline.
    d. Some estimators can also do predictions via predict().
    e. All steps but the last one in a pipeline are Transformers.
4. Which of the following is a key advantage of XGBoost over traditional gradient boosting methods?
    a. It uses a linear regression model instead of decision trees.
    b. It employs optimization techniques like parallel processing.
    c. XGBoost can only work with numerical features and requires all categorical variables to be dropped before training.
    d. XGBoost uses approximate histogram computation to speed up threshold finding.
5. Which of the following statements about neural networks are true?
    a. There is an input neuron for each sample in the data.
    b. Backpropagation computes gradients by propagating errors backward from the output layer to the input layer.
    c. Activation functions must be linear to allow the network to learn complex patterns.
    d. The forward pass involves computing predictions by passing input data through the network layers, applying weights, biases, and activation functions at each layer.
6. What are some advantages of fully sharded data parallelism (FSDP) over data parallelism (DP) in distributed learning?
    a. FSDP allows each GPU to train on different model architectures simultaneously, enabling ensemble learning within a single training run.
    b. FSDP eliminates the need for gradient synchronization between GPUs, making it faster than traditional data parallelism for all model sizes.

c. FSDP improves memory efficiency by avoiding the redundancy of storing identical copies of model parameters on every GPU.
d. FSDP reduces memory consumption per GPU by sharding model parameters, gradients, and optimizer states across devices, enabling training of larger models.