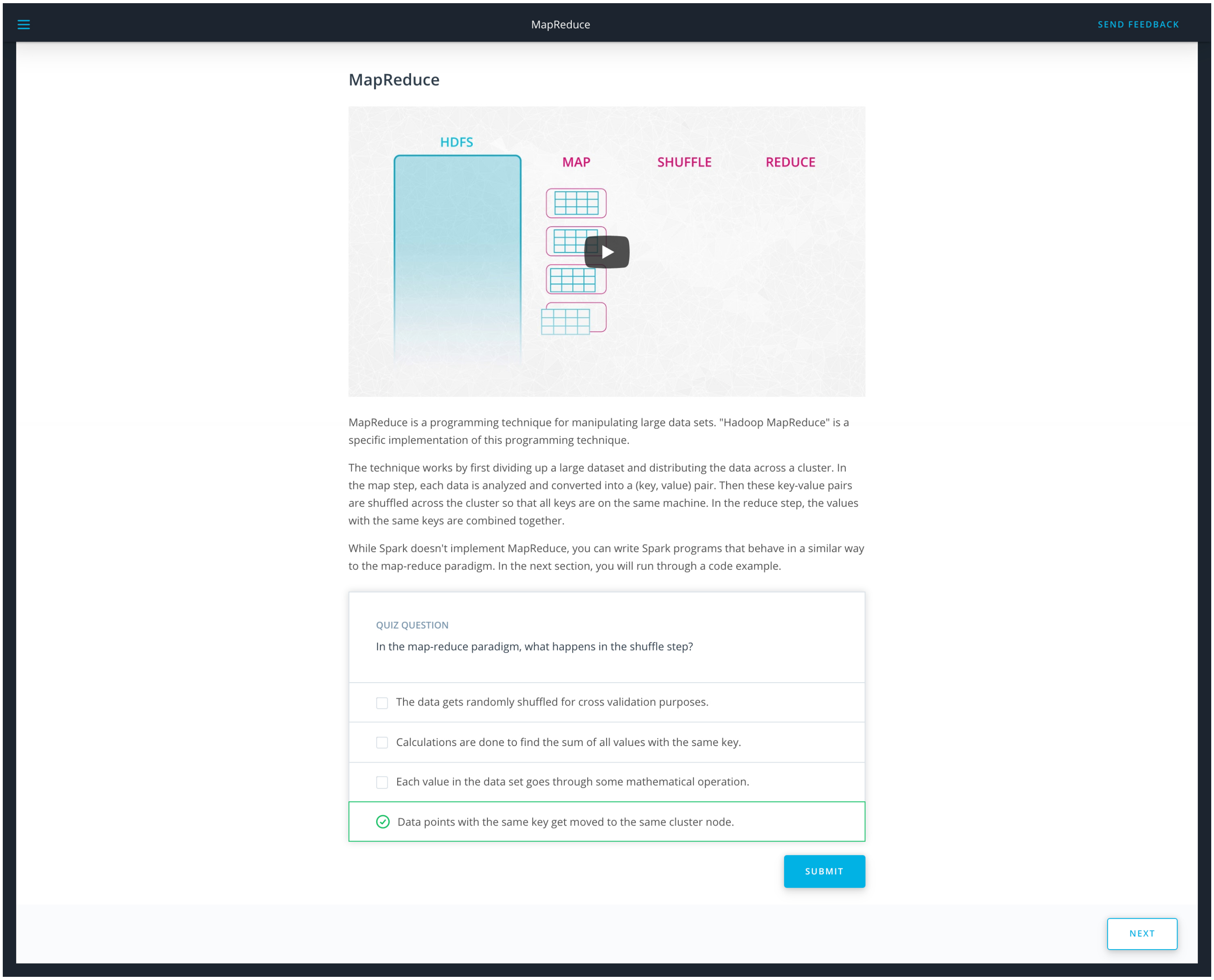

MapReduce

MapReduce is a programming technique for manipulating large data sets. "Hadoop MapReduce" is a specific implementation of this programming technique.

The technique works by first dividing up a large dataset and distributing the data across a cluster. In the map step, each data is analyzed and converted into a (key, value) pair. Then these key-value pairs are shuffled across the cluster so that all keys are on the same machine. In the reduce step, the values with the same keys are combined together.

While Spark doesn't implement MapReduce, you can write Spark programs that behave in a similar way to the map-reduce paradigm. In the next section, you will run through a code example.

QUIZ QUESTION

In the map-reduce paradigm, what happens in the shuffle step?

- □ The data gets randomly shuffled for cross validation purposes.
- □ Calculations are done to find the sum of all values with the same key.
- □ Each value in the data set goes through some mathematical operation.
- ✓ Data points with the same key get moved to the same cluster node.