School of Electrical Engineering & Computer Science
INFS3208 Cloud Computing

THE UNIVERSITY OF QUEENSLAND

Homework 8: Spark Framework

Correct answers are labeled red bold.
Some necessary explanations are labeled blue.

Q1: What are the features of Spark RDD?

A. In-memory computation
B. Lazy evaluation
C. Fault tolerance
D. All of the above

Q2: How many Spark Context can be active per JVM?

A. Two
B. One
C. Unlimited
(Only one SparkContext may be active per JVM. You must stop() the active SparkContext before creating a new one.)

Q3: Which of the following is not a transformation?

A. `flatMap(func)` (`flatMap(func)` is a transformation operation and it's similar to `map(func)`, but each input item can be mapped to 0 or more output items (so `func` should return a Seq rather than a single item).)
B. `map(func)`(`map(func)` is a transformation operation and returns a new distributed dataset formed by passing each element of the source through a function `func`)
C. `reduce(func)`(`reduce(func)` is an action operation and it aggregates the elements of the dataset using a function `func` (which takes two arguments and returns one). The function should be commutative and associative so that it can be computed correctly in parallel.)
D. `filter(func)` (`filter(func)` is a transformation and returns a new dataset formed by selecting those elements of the source on which `func` returns true.)

Q4: Which one of the following operations does NOT trigger an eager evaluation?

A. `take(n)` (`take(n)` is an action operation and returns an array with the first n elements of the dataset.)
B. `collect()` (`collect()` is an action and returns all the elements of the dataset as an array at the driver program. This is usually useful after a filter or other operation that returns a sufficiently small subset of the data.)
C. `count()` (`count()` is an action and returns the number of elements in the dataset.)
D. `join(otherDataset,[numPartitions])` (`join(otherDataset, [numPartitions])` is a transformation operation and won't trigger the evaluation.)

Q5: Which of the following commands will NOT generate a shuffle of data from each executor across the cluster?

A. `collect()`

B. `map(func)`

C. `repartition(numPartitions)`

D. `distinct([numPartitions]))`

(`map()` transformation is narrow and does not trigger a shuffle. The other options, such as `repartition()`, `distinct()` and `distinct()`, typically cause data shuffling across the cluster.)