RESILIENT DISTRIBUTED DATASETS

(1/1 point)

Which of the following is not a property of RDDs?

Top of Form

They can be changed after they are constructed They can be changed after they are constructed - correctThey can be created by transformations applied to existing RDDsThey enable parallel operations on collections of distributed dataThey track lineage information to enable efficient recomputation of lost data

Bottom of Form

**EXPLANATION**

RDDs cannot be changed once they are created - they are immutable. You can create RDDs by applying transformations to existing RDDs and Spark automatically tracks how you create and manipulate RDDs (their lineage) so that it can reconstruct any data that is lost due to slow or failed machine. Operations on RDDs are performed in parallel.

SPARK TRANSFORMATIONS

(1/1 point)

Which of the following are properties of Spark transformations?

Top of Form

They are not computed right away, They are like a recipe for creating a result, - correct

They are not computed right away They are computed right awayThey are vulnerable to machine failuresThey are like a recipe for creating a result

Bottom of Form

Note: Make sure you select all of the correct options—there may be more than one!

**EXPLANATION**

Spark Transformations use lazy evaluation, which means they are not immediately executed. Instead they can be thought of as a recipe for creating a result from an input dataset.

SPARK ACTIONS

(1/1 point)

Which of the following is not a property of Spark Actions?

Top of Form

They cause Spark to execute the recipe to transform the source dataThey are the primary mechanism for getting results out of SparkThey are lazily evaluated They are lazily evaluated - correctThe results are returned to the driver

Bottom of Form

**EXPLANATION**

Spark Actions are the mechanism for causing Spark to apply the specified set of transformations to the source data. They are the way that you extract the results out of Spark at the driver.

SPARK PROGRAM LIFECYCLE

(1/1 point)

Which of the following are part of a Spark program's lifecycle?

Top of Form

RDDs that are reused may be cached, Actions cause parallel computation to be immediately executed, Transformations lazily create new RDDs, - correct

RDDs that are reused may be cached Transformations cause parallel computation to be immediately executedActions cause parallel computation to be immediately executed Transformations lazily create new RDDs Actions create recipes for peforming parallel computation on datasets

Bottom of Form

Note: Make sure you select all of the correct options—there may be more than one!

**EXPLANATION**

Transformations specify how to perform parallel computation in a lazily evaluated manner. Actions cause the transformations to be executed. If you plan to reuse an RDD, you should cache it.

PYSPARK SHARED VARIABLES

(2/2 points)

In iterative or repeated computations, broadcast variables avoid the problem of sending the same data to workers:

Top of Form

FalseTrue True - correct

Bottom of Form

**EXPLANATION**

Broadcast variables are an efficient way of sending data once that would otherwise be sent multiple times automatically in closures.

Accumulators can be used by Spark workers to efficiently read values during distributed computations:

Top of Form

False False - correctTrue

Bottom of Form

**EXPLANATION**

Accumulators can only be written by workers and read by the driver program.