Bookmarks

- Week 1 - Big Data and Data Science
- Week 2 - Performing Data Science
- **Week 3 - Programming with Resilient Distributed Datasets**

Lecture 3: Apache Spark Resilient Distributed Datasets
Quizzes

Lab3a - RDD Tutorial
Lab due Sep 13, 2016 at 04:30 IST

Lab 3b - Text Analysis and Entity Resolution
Lab due Sep 13, 2016 at 04:30 IST

Lab 3b Quiz Questions
Quizzes

Bookmark

Resilient Distributed Datasets (RDDs)

BERCS1102016-V001700

0.00 / 16.17 1.0x

Start of transcript. Skip to the end.

SPEAKER: We can create an RDD from a Python collection,

a list.

So here we have an example with the list data

consisting of the elements 1, 2, 3, 4, and 5.

And we create an RDD by calling sc.parallelize,

Download video Download transcript .srt

Note that when using sortByKey(), the ordering of values for a key is non-deterministic.

Spark RDD Transformations

(1/1 point)
Which of the following are properties of Spark RDD transformations?

- [x] They are not computed right away ✔
- [ ] They are computed right away
- [ ] They are vulnerable to machine failures
- [x] They are like a recipe for creating a result ✔

✔

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

EXPLANATION

Spark RDD 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 RDD Actions

(1/1 point)

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

- They cause Spark to execute the recipe to transform the source data
- They are the primary mechanism for getting results out of Spark
- They are lazily evaluated ✔
- The results are returned to the driver

EXPLANATION

Spark RDD 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.