BerkeleyX: CS110x Big Data Analysis with Apache Spark

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

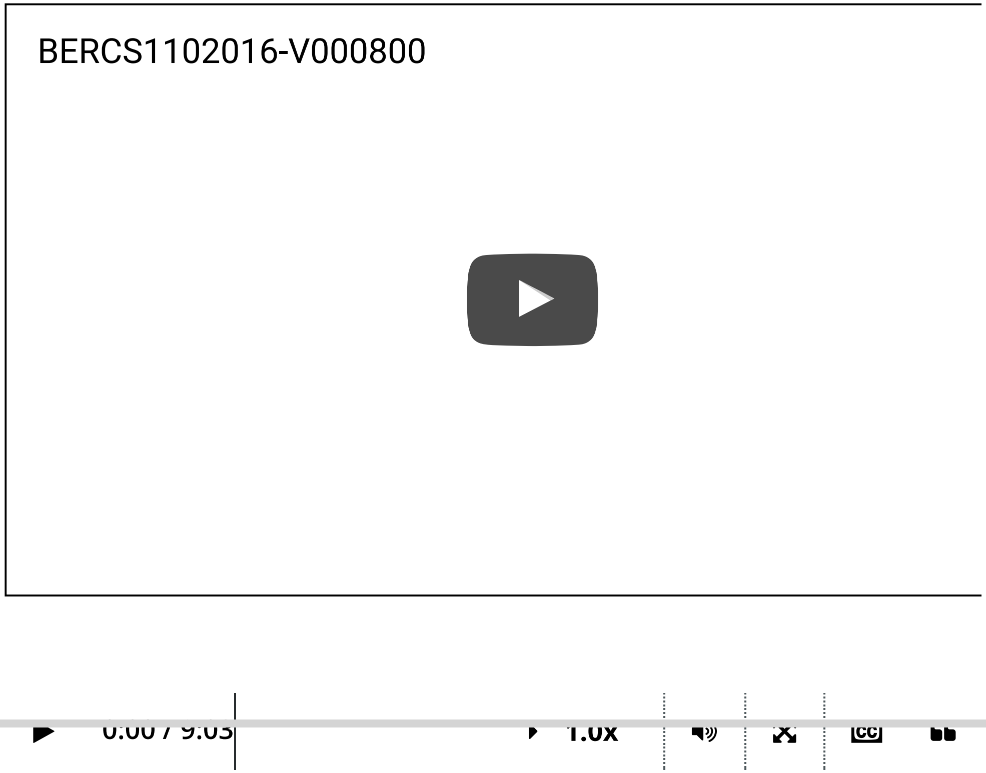
Lab 3b - Text Analysis and Entity Resolution
Lab due Sep 13, 2016 at 04:30 IST

Lab 3b Quiz Questions
Quizzes

Week 3 - Programming with Resilient Distributed Datasets > Lecture 3: Apache Spark Resilient Distributed Datasets > DataFrames and Resilient Distributed Datasets

Bookmark

DataFrames and Resilient Distributed Datasets



Start of transcript. Skip to the end.

SPEAKER: In this lecture, we'll look at Resilient Distributed

Datasets, we'll look at how to create an RDD,

we'll look at Spark's RDD transformations and actions,

we'll look at the Spark RDD programming model,

and we'll look at Spark shared variables

Download video Download transcript .srt

Apache Spark Resilient Distributed Dataset References

For more information about Apache Spark's RDDs, you should refer to the online Spark Documentation. The documentation includes screencasts, training materials, and hands-on exercises. The Spark Programming Guide is another good starting point, and the pySpark RDD API documentation is a great reference to use when writing Spark programs using RDDs.

Resilient Distributed Datasets

(1/1 point)

Which of the following is not a property of RDDs?

● They can be changed after they are constructed ✓

○ They can be created by transformations applied to existing RDDs

○ They enable parallel operations on collections of distributed data

○ They track lineage information to enable efficient recomputation of lost data

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.

These are the links in this lecture:

- Spark Documentation
- Spark Programming Guide
- pySpark RDD API documentation
- Python Documentation
- Download Python Documentation

RDDs versus DataFrames (I)

(1/1 point)

You should use DataFrames when you have data that is

● semi-structured or structured ✓

○ freeform or unstructured

○ video or audio

EXPLANATION

DataFrames are the ideal choice when your data is semi-structured or structured.

RDDs versus DataFrames (II)

(1/1 point)

You should use RDDs when you have data that

○ benefits from Project Tungsten's memory management

● needs low-level transformations or actions ✓

○ requires high-level transformations or actions

EXPLANATION

RDDs are the ideal choice when you need to perform low-level transformations or actions on your data.