# Master

databricks


## Lesson 21 - PySpark

## *Using RDDs*

**by**

**Bryan Cafferky**

***Data Enabler***

***from YouTube Channel***

# Where Are We Going?

- ✓ What is an RDD?
- ✓ Lazy Evaluation
- ✓ Transformations
- ✓ Actions

# What is an RDD?

- ✓ Resilient Distributed Dataset
- ✓ Fault-tolerant collection of elements that can be operated on in parallel.
- ✓ RDDs are Immutable
- ✓ Fundamental Spark Data Structure

# Lazy Evaluation

- ✓ Spark will only do something when forced to
- ✓ Transformations are not done until an Action is called
- ✓ Transformations create a new RDD from an existing RDD
- ✓ Actions return the results to the Driver

# What is a Transformation?

- ✓ Applies logic to the dataset to change it
- ✓ map() – Pass each element through a function
- ✓ filter() – Select elements to retain
- ✓ sample() – Return a subset of the dataset.

# What is an Action?

- ✓ Executes pending transformations
- ✓ Returns results to the driver
- ✓ count() – Return the number of elements
- ✓ reduce() – Return an aggregation
- ✓ collect() – Return the results to the driver (caution)
- ✓ take(n) – Returns n rows to the driver.

Demonstration

# Wrapping Up

- ✓ What is an RDD?
- ✓ Lazy Evaluation
- ✓ Transformations
- ✓ Actions

# Thank You!