Serialization

Objective

Fix the "task not serializable" problems

Understand how Spark serializes transformations for executors

Serialization Problems

Applicable to lambdas (RDDs and Datasets only)

Task serialization needed

- driver needs to send lambdas to the executor JVMs for them to execute
- every reference the lambdas make needs to be serialized as well
- if reference is a field of a class/object, it needs to be serialized
- not everything is serializable

Technique 1: make classes/objects Serializable

- Java marker interface for serializable instances
- easy to use
- large overhead

Technique 2: capture local values

- harder to use
- optimal

Spark rocks