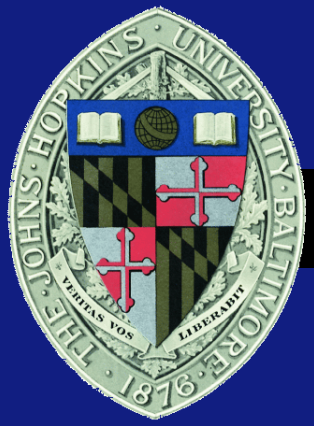
Lecture 16.2
Spark Map Reduce

EN 600.320/420

Instructor: Randal Burns

5 April 2017



Department of Computer Science, *Johns Hopkins University*

Map/Reduce in Spark

- The following steps would be equivalent:
  - spark.textfile(...).flatMap(...).reduceByKey(...).save()
  - Doesn't use RDD pipelining. flatMap produces a sequence.
  - Doesn't use memory abstraction

Comments about Transformations

- map() is one-to-one consistent w/ scala semantics
  - flatMap is many-to-one like Map in M/R
  - mapValues: does not transform key (important for partitioning)
- Most transformations are one-to-one
  - Important for partitioning
- Others are not one-to-one: called “wide-dependencies”
  - Bad for partitioning
  - Will discuss later