Distributed Systems Project, Spring 2016

Assignment 2
Assignment: Spark

• Use Spark to analyze large data sets

• Goals:
  • Get familiar with Spark
  • Understand how to work with large data sets
Two Tasks

• Task 1:
  • Calculate statistics from a large set of numbers
  • Two sub-tasks calculation, one sub-task explaining

• Task 2:
  • Matrix multiplication with a very large matrix
Environment

• We have set up Spark on Ukko cluster
• Assignment sheet contains instructions
• Help with Spark during Q&A
• Help with matrices if needed
Next Steps

• 2.2: Q&A and Spark tutorial, matrix math if needed
• 4.2., 9.2., 11.2. Normal Q&A sessions
• 16.2: Deadline (extended)