Distributed Systems Project, Spring 2015

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

• 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

• 22.2: Q&A and Spark tutorial, matrix math if needed
• 27.1 and 29.1: Normal Q&A sessions
• 3.2 and 5.2: Normal Q&A
• 10.2: Deadline (extended)