UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

INF-2202 (Fall 2016)
Assignment 3

PageRank using Spark on Amazon Web Services

Tim A. Teige & Lars Ailo Bongo
11.10.2016

Overview

- Your task is to implement PageRank using Spark and Amazon Web Services
- PageRank - Used to rank web pages for search engines
- Spark - Open source cluster computing framework
- You can either use Python or Scala to do this assignment
- **Deadline 07.11**

PageRank

- Assigns a rank based on number of links to and from.

- Algorithms maintains two data sets
  - (pageId, link list)
  - (pageId, rank)

- Iterative algorithm

1. Default initial value of the rank is 1.0.
2. Each iteration page p send a contribution of rank(p)/numNeighbours(p) to its neighbours(pages which it links to)
3. Set each page’s rank to 0.15 + 0.85 * contributionReceived

- The last two steps are repeated for a number of iterations, typical value is 10.

Spark

- Computing framework for clusters
- You will run Spark on the AWS clusters.
- You will use ssh to log in to the cluster and runs jobs (specifics detailed in assignment readme)
- Lecture on Spark at Thursday 13.10
- Next group session (18.10) will be used as a walkthrough on AWS and using Spark

Dataset

- The dataset that will be used is the **Common Crawl Corpus**.
- Openly available from inside the S3 service from amazon or at http://commoncrawl.org/the-data/get-started/
- Total datasize is too large compared to the available funds on your Amazon account, so use a subset of the dataset.

Amazon Web Service Account

- You must create an AWS account in order to do this assignment
- There is no need to register a credit card account
- Create an account at https://aws.amazon.com/education/awseducate/apply/
- You must use your university email and apply as a student.
- This will provide you with 75$ to use when actually running your program on the cluster.

Requirements

- Implement PageRank using Spark
- Measure Performance
- Report describing the implementation, design and a performance evaluation

Github Workflow

- An invitation link to the assignment will be sent out on the mailinglist
- Work in your own private repository
- Delivery in the github repository

Grading

- **APPROVED** or **NOT APPROVED**

- Evaluation based on the implementation and the report

Disclaimer

- Please do not publicize or share your solution or codes anywhere without our permission

- Please refrain yourself to copy other students code(s).

- On the contrary, group discussions and brainstorming for ideas are strongly encouraged