Enabling data-intensive scientific workflows with Apache Spark

Supervisors: Ola Spjuth (ola.spjuth@farmbio.uu.se), Marco Capuccini (marco.capuccini@it.uu.se)
Institution: Department of Pharmaceutical Biosciences, Uppsala University

High-throughput methods in various scientific fields produced massive datasets in the past decades, and using Big Data analytics frameworks such as Apache Spark (http://spark.apache.org) is a natural choice to enable data-intensive analysis. In scientific applications, many tools are highly optimized to resemble or detect some phenomena that occur in a certain system. Hence, the effort of reimplementing scientific tools in Apache Spark can seldom be sustained by research groups. As part of our research activity in the PhenoMeNal project (http://phenomenal-h2020.eu) we implemented MaRe (https://github.com/mcapuccini/MaRe) a MapReduce-oriented parallelization library that allows to parallelize any existing scientific tool by leveraging on Apache Spark and Docker (https://www.docker.com). In this project we aim to integrate MaRe with a workflow system such as NextFlow (https://www.nextflow.io/) or Spotify Luigi (https://github.com/spotify/luigi). This will enable to run fully automated in-silico pipelines over large-scale datasets, while integrating with the Apache Spark ecosystems. Furthermore, this will open interesting opportunities such as streaming and interactive analysis for scientific pipelines.

Desired skills

- Fluent at coding in Java or Python (knowledge of Scala and Groovy is also relevant)
- Good understanding of the Big Data ecosystem (Apache Spark and Hadoop)
- Good understanding of containerization engines (Docker)

Project plan

- Run some simple applications in MaRe, NextFlow and Luigi
- Implement MaRe integration in NextFlow or Luigi
- Run a benchmark on the SNIC Science Cloud