Big Data Assignment 3: *Spark RDDs*

Meet Spark and gain in-depth knowledge of the Resilient Distributed Dataset (RDD). The first learning objective is to acquire the competences to carry out basic data tasks with Spark, and contrast the user experience to that of using the Map Reduce framework. A secondary objective is to learn to work with Zeppelin Notebooks, that will save you time later on in the course. The third objective is to improve your understanding of the Spark execution model, by looking into detail into Spark jobs, tasks, partitioning and more.

General instructions

In this assignment, you start working with Apache Zeppelin, a web-based notebook to work with Spark interactively (something similar to Jupyter notebooks which you might know). The exercises can be found in the attachments. How to open the Zeppelin notebooks is written in the next section.

The deadline for this assignment is **March 19th, 23:59**. Please hand in the Exercises notebook (`.zpln`) after filling it in.

After handing in the assignment, you can go to Brightspace Quizzes to test if you understood the assignment. Making the quiz is voluntary, but recommendable, as you can check your answers from the assignment there. If you have any questions (you find out your answer is incorrect and you do not know why), do not hesistate to come discuss it with us in the practicum (or send us an email / contact us in the matrix room, if going to campus is not an option for you).

Open the Zeppelin notebooks

Create the docker container for this assignment (it's another one than the one with access to our cluster) with the following docker command:

```
docker create --name big-data -it -p 8080:8080 -p 9001:9001 -p 4040:4040 \
rubigdata/course
```

(Ports 8080 and 4040 are used by Spark, we configured Zeppelin to appear on port 9001.)

Start the container:

```
docker start big-data
```

Now, connect to the Zeppelin UI (localhost:9001) from your internet browser; giving it a little time to start all the services inside the container.

Import the two notebooks you can find in the attachment, and start exploring Spark! *You find a screenshot for importing a notebook below; type the desired name, click on Select JSON File, and select the corresponding `.zpln` file that you downloaded from here.*

!Warning! If you work on a computer in one of the terminal rooms in Huygens, be aware that docker is reset every once in a while, which means you lose your existing containers. So, if you make any changes to a notebook and wish to keep those, make sure to `export this note (zpln)` (by clicking on the icon in the menu bar in Zeppelin).