Ciências ULisboa | Informática

Cloud Computing

**Project 3**
Big data with Hadoop and Spark

---

# 1 Introduction

Logs are being made by many applications for many different purposes. These logs explicitly contain a set of data per event. But the logs also possess valuable information that comes from the correlation of information from many events.
The goal of this project is to analyze a log file and extract more value from it than the mere event data.
The logs to be analyzed contain communication information with the following fields:
Source IP and port, destination IP and port, protocol, number of packets, size, flags, start time, duration, end time, sensor that acquired the data and a label. An example is show bellow.

```
sIP,dIP,sPort,dPort,protocol,packets,bytes,flags,sTime,duration,eTime,sensor,label
192.168.0.94,192.168.0.200,34618,80,6,55,2989,FSPA,1416846921.828,0.027,1416846921.855,S0,NORMAL
```

The specific goals of the project are to gather information that helps to answer the following questions:

1. Who sends more information: total and average per day?
2. With whom each source communicates and how much information is exchanged between pairs (source, destination)?
3. How frequently a pair of machines communicate (flows per day)?
4. Who communicates with more destinations?
5. Who sends and who receives more information?
6. Is anyone acting as a gateway?

To attain these goals, in the first phase of the project the solution should be based on Hadoop MapReduce. While on the second phase the challenge will be tackled with Spark.

# 2 Development and testing the system

The development can take place in any machine, but the usual machine, t5, is available for development and testing.

The system can be tested using the dataset SiLK-LBNL-05-nonscan.tar.gz (containing multiple files, in different directories) available in https://tools.netsa.cert.org/silk/referencedata.html. To make things easy for students, a csv version of this dataset (all files in a single dataset) is available on the moodle page of the course, together with this project description. The dataset was also uploaded to t5 with the path: `/root/project3-dataset.csv.zip`.

# 3 Report and delivery

Each group should write a report covering the solutions and difficulties found in the development of the project. Explain how the setup and testing was made and prepare a final discussion supporting the main conclusions.
All the files developed by the students and all configurations should also be delivered as a zip file.
This work should be completed until the 4th of June, 2017. The report and zip file should be delivered to both lecturers by using the following addresses: mcalha@ciencias.ulisboa.pt and anbessani@ciencias.ulisboa.pt.