GCD Week 6 – Spark

Kevin Jetten

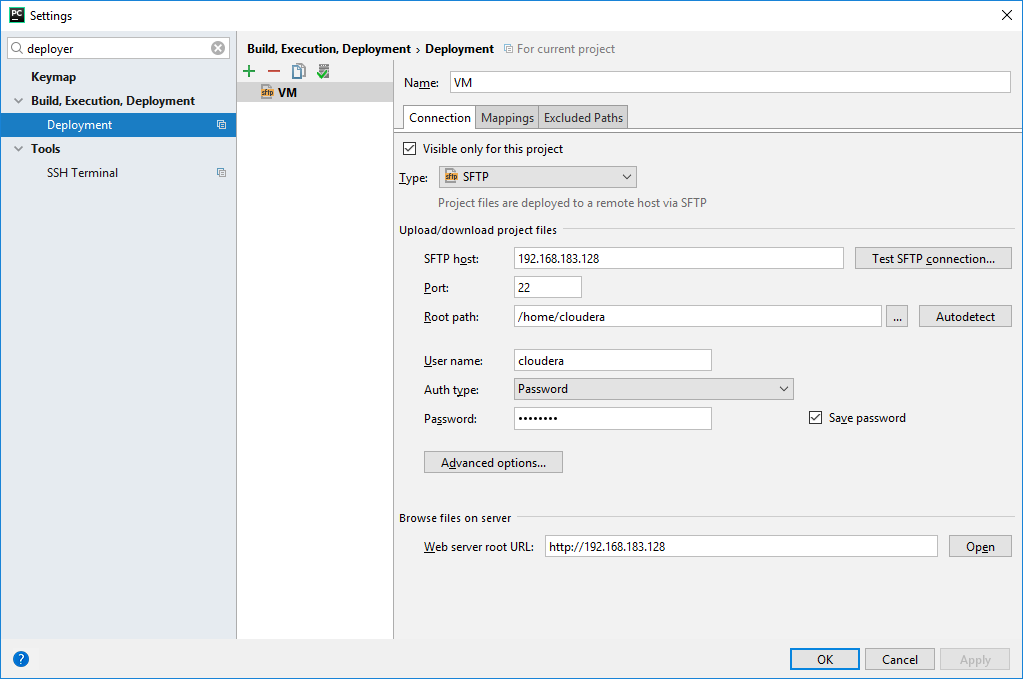
i318843

Inhoud

[Activity 1: Analyse a weblog 3](#_Toc496277779)

# Preparation

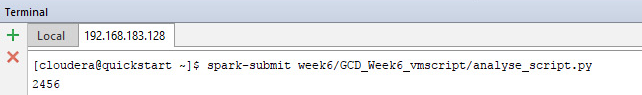
To avoid having to develop python scripts on the VM, I added the VM and spark as an interpreter to my Pycharm installed on my normal system.

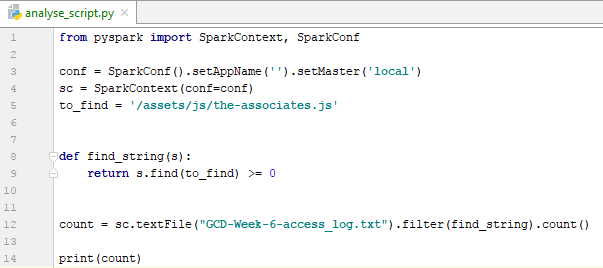


# Activity 1: Analyze a weblog

## Activity 2.1: how many hits are there to page ‘/assets/js/the-associates.js’?

**The answer is: 2456**

****

****

## Activity 2.2: how many hits are there from ip address 10.99.99.186?

**The answer is: 6**

To avoid adding nearly the same screenshot again I won’t post this code because it is the same as the one above except I changed the value of ‘to\_find’ to ‘10.99.99.186’

# Activity 2.3: Find the most popular file on the website, that is, whose path occurs most often in access\_log. The output should be: the file’s path and the number of times it occurs in the log.