***Kaggle Data Science Platform***

***User Guide***

***Version 0.3 / 26-06-2017***

[1 Introduction 2](#_Toc486214536)

[1.1 About the platform 2](#_Toc486214537)

[1.2 Kaggle teams 2](#_Toc486214538)

[1.3 Rules 2](#_Toc486214539)

[2 Access to the Platform 2](#_Toc486214540)

[2.1 SSH access 3](#_Toc486214541)

[2.2 Server mode access 4](#_Toc486214542)

[2.2.1 Zeppelin server: notebooks for data science (Python, R, Spark, Hive, bash) http://52.57.83.1:8080/ 4](#_Toc486214543)

[2.2.2 R Studio server: http://52.57.83.1:8787/ 5](#_Toc486214544)

[2.2.3 Kaggle data location: /accenture-kaggle-datascience 6](#_Toc486214545)

[2.2.4 Spark Server: http://52.57.83.1:4040 7](#_Toc486214546)

[2.2.5 Git repository: /data/opt/git 7](#_Toc486214547)

[2.2.6 KNIME program for data preparation: /data/opt/knime 7](#_Toc486214548)

[2.2.7 JupyterHub server: traditional IPython style data science notebooks http://52.57.83.1:8000/ 7](#_Toc486214549)

[2.2.8 Work in progress 7](#_Toc486214550)

# Introduction

This is the user guide for Kaggle Data Science platform.

## About the platform

Kaggle Data Science platform consists of:

1. 1 x Amazon EC2 computing instance of type t2.large - medium size (4 CPU, 8 GB RAM) runs on an Amazon Ubuntu Linux 16.04 server; fixed IP allocated for this cloud computer (please note this IP can change): **52.57.83.1 /** .
2. 1 x Amazon EMR (Hadoop-as-a-Service) cluster consisting of 1 x core node + 2 x task instances
3. 1 x Amazon S3 bucket for data storage.

To access the data science platform, you need to obtain a Linux user account which has same value as your Accenture enterprise ID. Participants in Kaggle competition that are members of Accenture NL Kaggle team can obtain a user account by sending a request to [r.serban@accenture.com](mailto:r.serban@accenture.com).

## Kaggle teams

This is the teams structure:

1. Team 1: Sihan Ding, Radu Serban, Stefan van Eijk and Sarah Parinussa
2. Team 2: Niklas Andersson, Luuk Figdor, Arjan Groen, Martijn de Graaff and Paulien Koeleman
3. Team 3: Laury van Bedaf, Olja Ljubicic, Theun van Vliet and Wouter Imanse

## Rules

1. You can use the data platform during agreed work hours and should accept that being a self-managed platform it might have issues. We are working to resolve them, and your input and suggestion is appreciated.
   * **Usage times**: business days (Mon-Fri) 08:00 .. 22:00. On request from users to email address [r.serban@accenture.com](mailto:r.serban@accenture.com) the platform can be started outside these hours, subject to agreement/approval with Paulien Koeleman.
2. Do not copy and share confidential data onto the platform.
3. Do not remove nor copy data that does not belong to you from the platform.
4. Respect the work of your colleagues in another team. Some of the tools and services are not (yet) configured to provide a complete separation of access and control of access restricted only to the resources of your team. Your help and understanding in working only with the resources expected and shared for your team (unless otherwise communicated) is appreciated.
5. Send any issues, comments and suggestions you might have as email to [r.serban@accenture.com](mailto:r.serban@accenture.com). Please include [Kaggle platform] as part of your subject line.

# Access to the Platform

You can access the data science platform:

1. In console mode: you can run own R, Python scripts/commands in a Linux session;
2. In server mode, from your web browser: using Zeppelin server, RStudio server;
3. In graphical mode: start an X server session by using MobaXTerm and run graphical programs remotely, e.g. Eclipse, KNIME, Anaconda IDE.

## SSH access

Pre-requisites: you need to have Open SSH client (putty, MobaXTerm portable home edition: <http://mobaxterm.mobatek.net/download.html> )

To access data science platform run the following command from command line:

* Open your SSH client program (putty, MobaXTerm)
* vim ~/.ssh/config

Host datascience.accenture.com

Hostname=52.57.83.1

User=<your Accenture enterprise ID, e.g. r.serban>

FForwardAgent yes

ForwardX11 yes

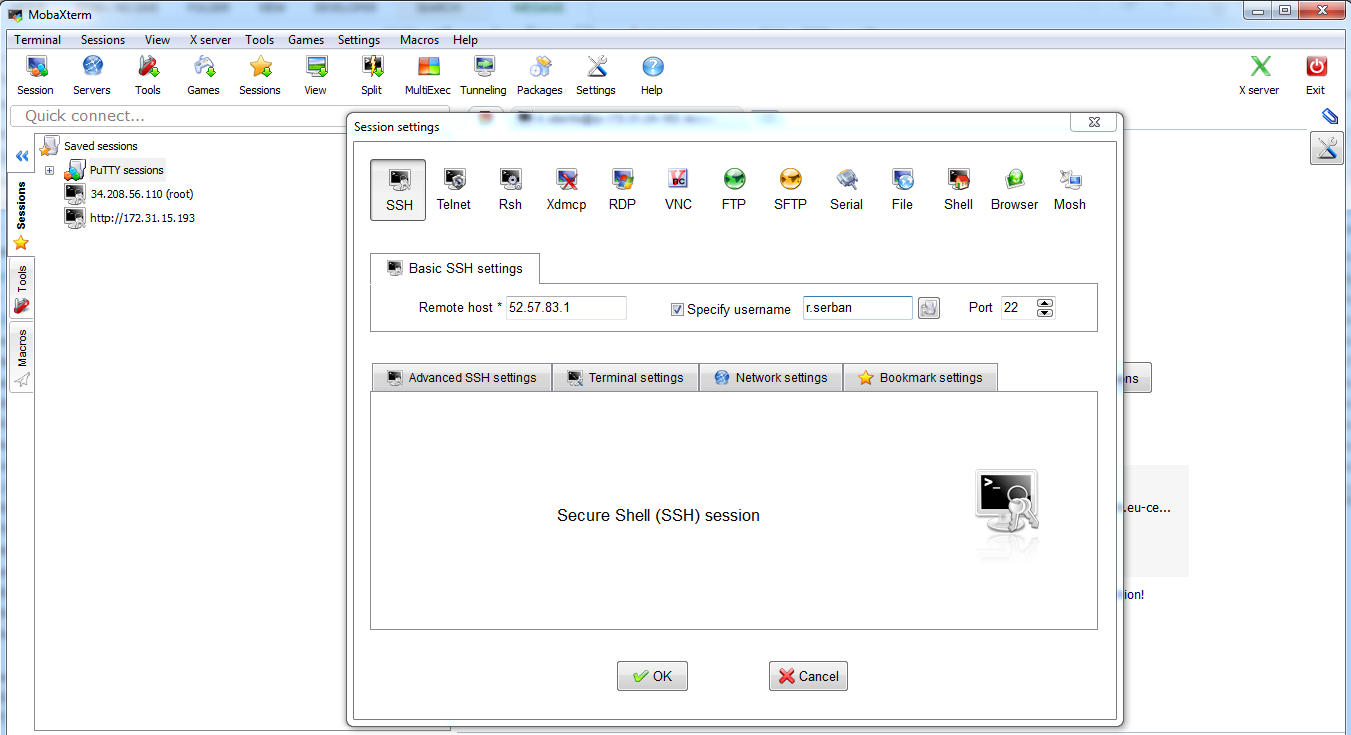
* ssh -X -o ForwardX11=yes <your Accenture enterprise ID, e.g. r.serban>@ec2-35-158-181-155.eu-central-1.compute.amazonaws.com
* ssh <your Accenture enterprise ID>@52.57.83.1

As password: <password/code-received-by-SMS or follow-up email>

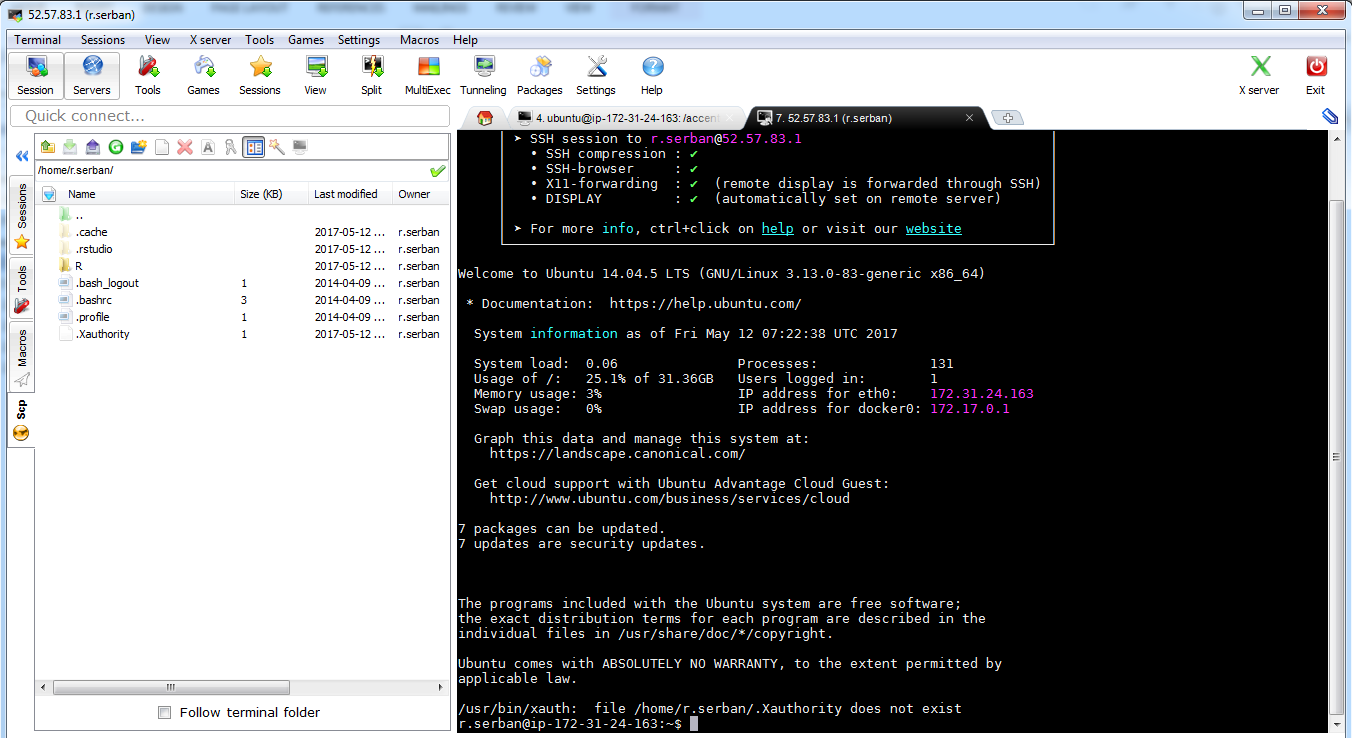
Once SSH connection is established, you have access to data science platform in console mode.

Your home folder is in /home/<your Accenture enterprise ID>

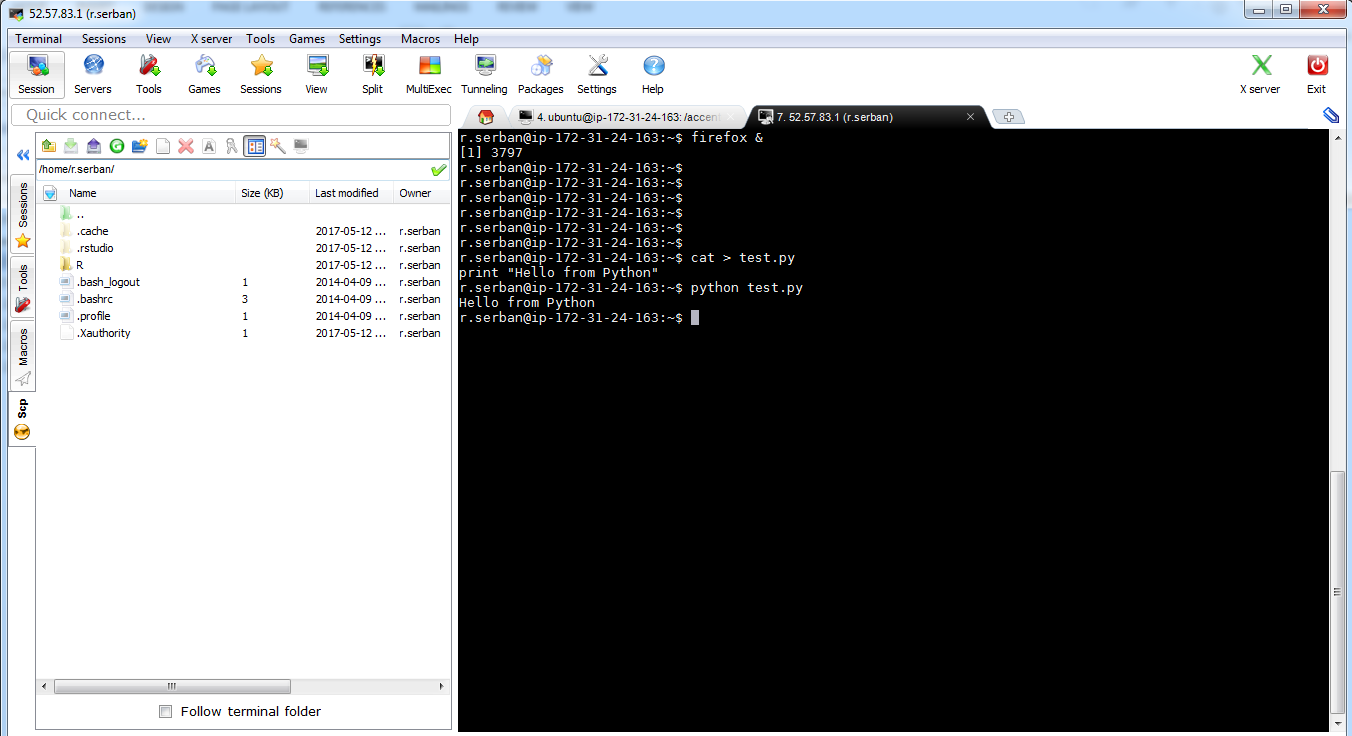
Note: you can change your password by running `passwd` command.



Once logged in, you can execute Linux commands.



You can execute your own Python/R scripts, and/or start graphical programs

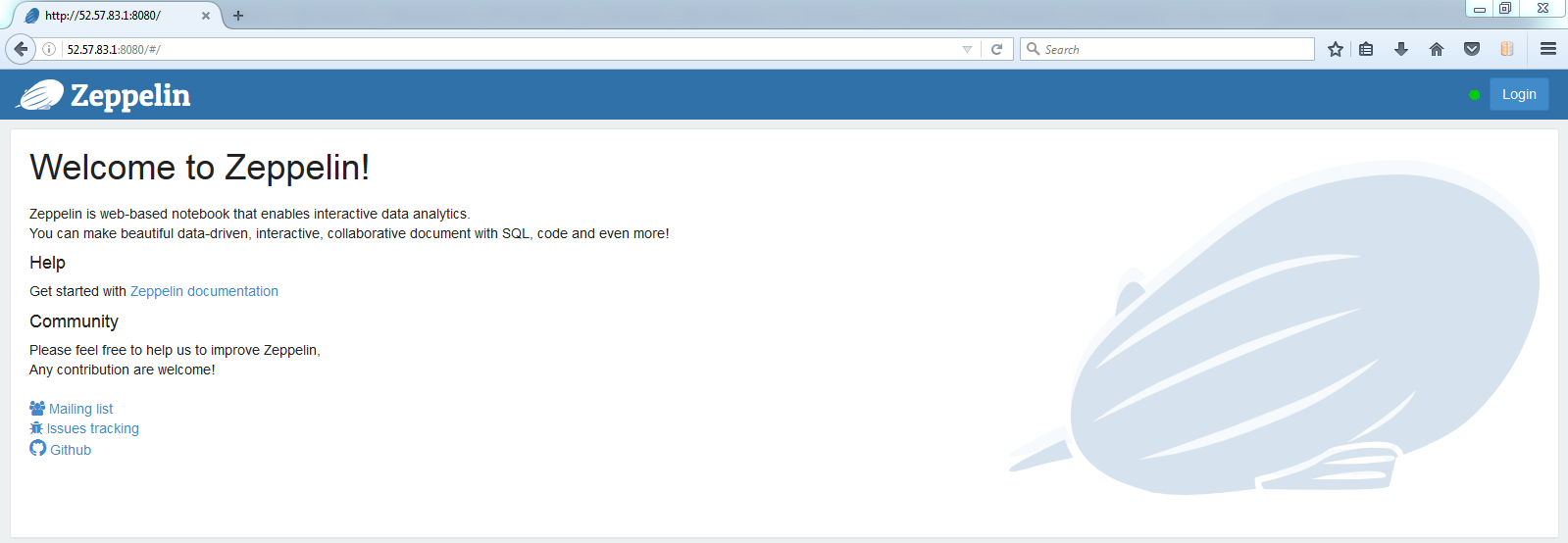


## Server mode access

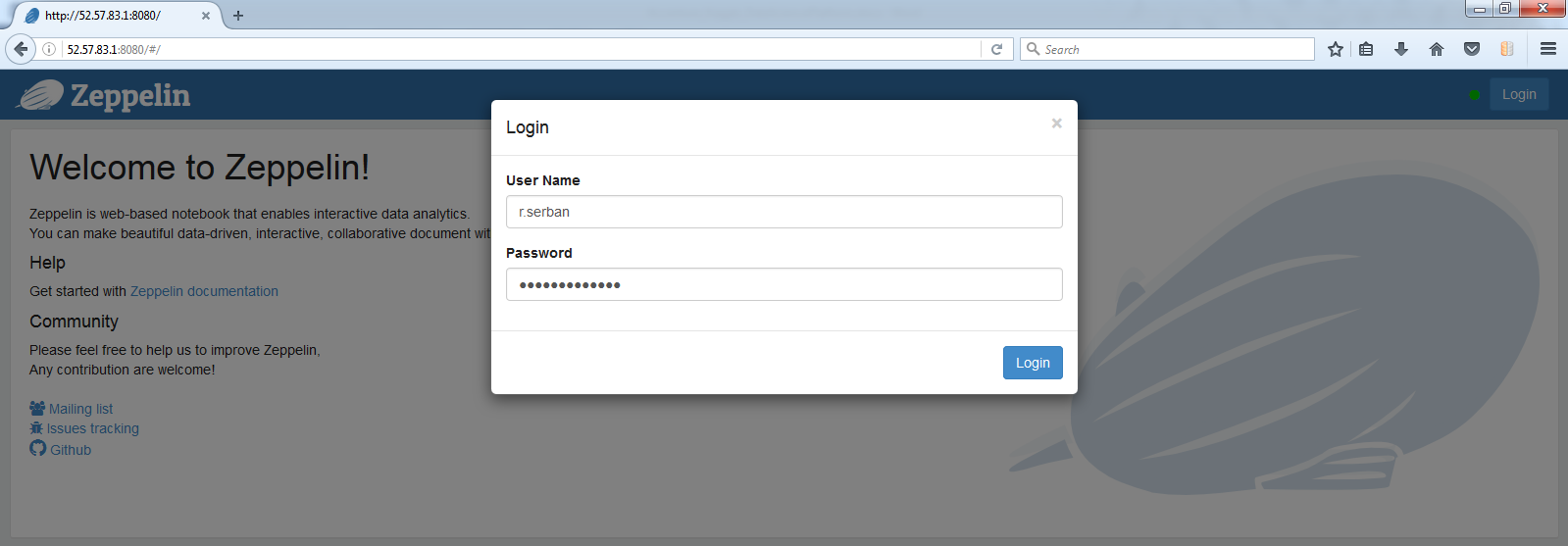
Open in your browser the following URLs:

### Zeppelin server: notebooks for data science (Python, R, Spark, Hive, bash) <http://52.57.83.1:8080/>

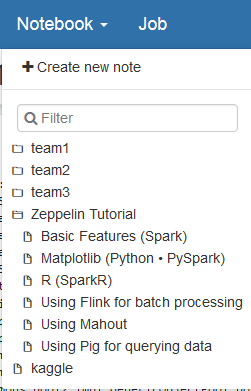
* + Zeppelin Server URL is <http://52.57.83.1:8080/>



* + Click “Login” and authenticate with your Linux account user ID and password:

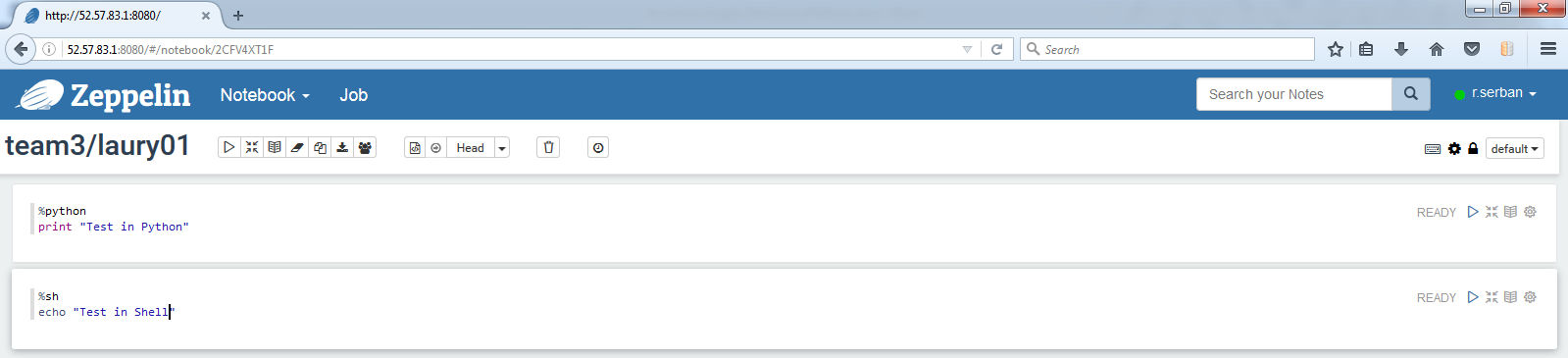


* + In here you can define your own R and Python notebooks. Each of you can create one or more notebooks, which are stored in folder /accenture-kaggle-datascience/NOTEBOOK/

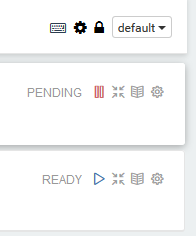


Even though you can create notebooks in other locations, please respect naming convention for each notebook, as team<X>/<FirstName lowercase>\_<NotebookNr>[\_<OptionalDescription “-“ for spaces>], e.g. for Radu team1/Radu\_01, for Laury team3/Laury\_01, team3/Laury\_02, etc

* + You can input your notebooks. Each section can have a different interpreter and contain both code and documentation



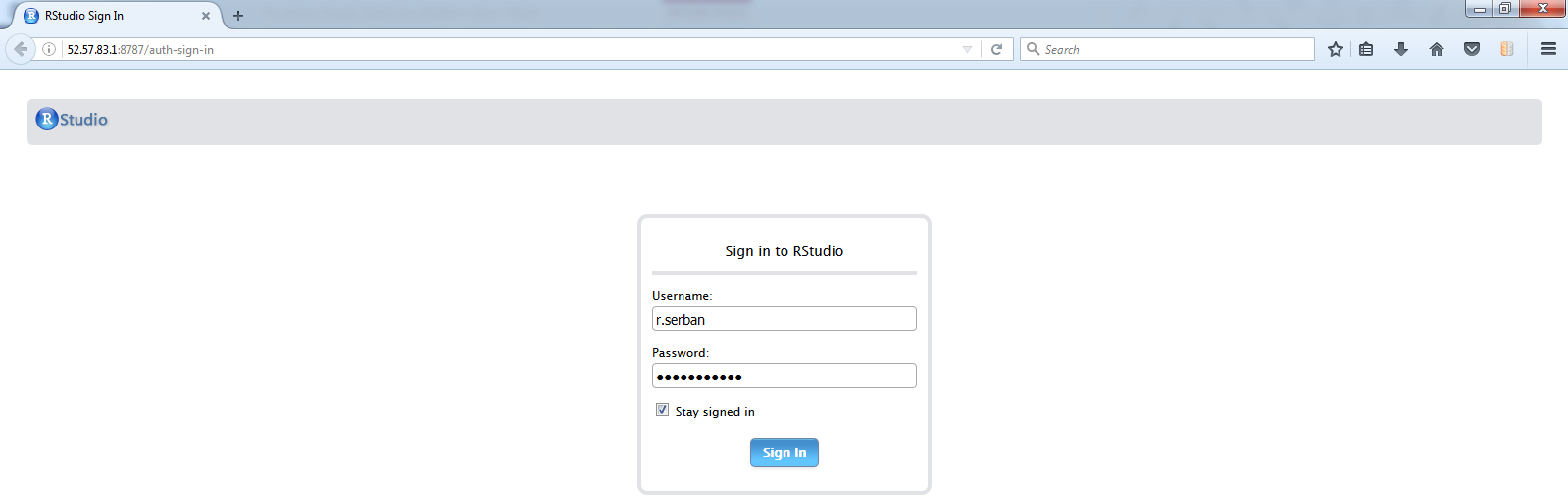
* + You can change between edit, execution and visualization mode and run code and/or visualize results from the notebooks



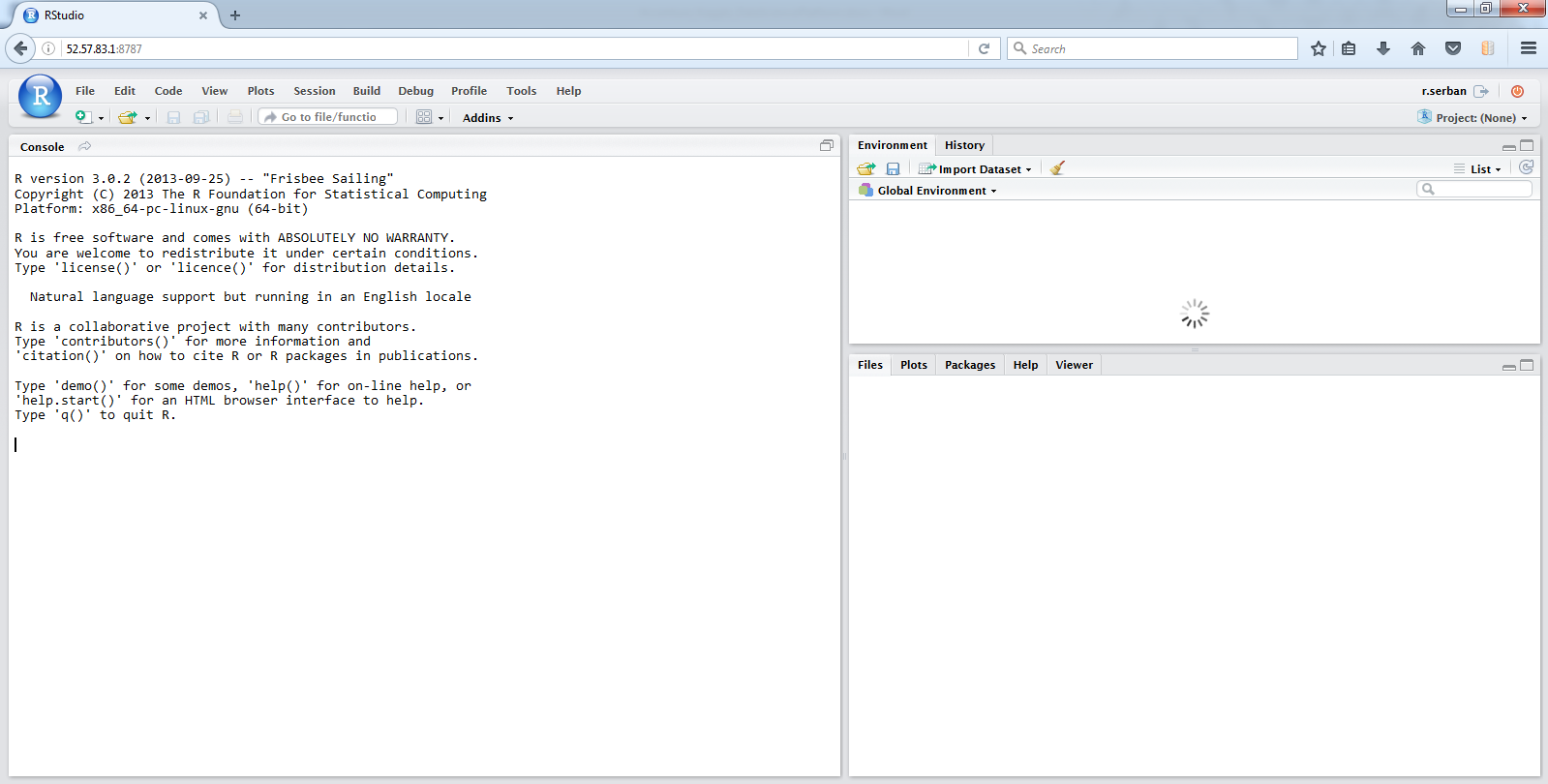
### R Studio server: <http://52.57.83.1:8787/>

* + RStudio Server URL: <http://52.57.83.1:8787/>

You authenticate with your Linux account user ID and password – the same as for SSH



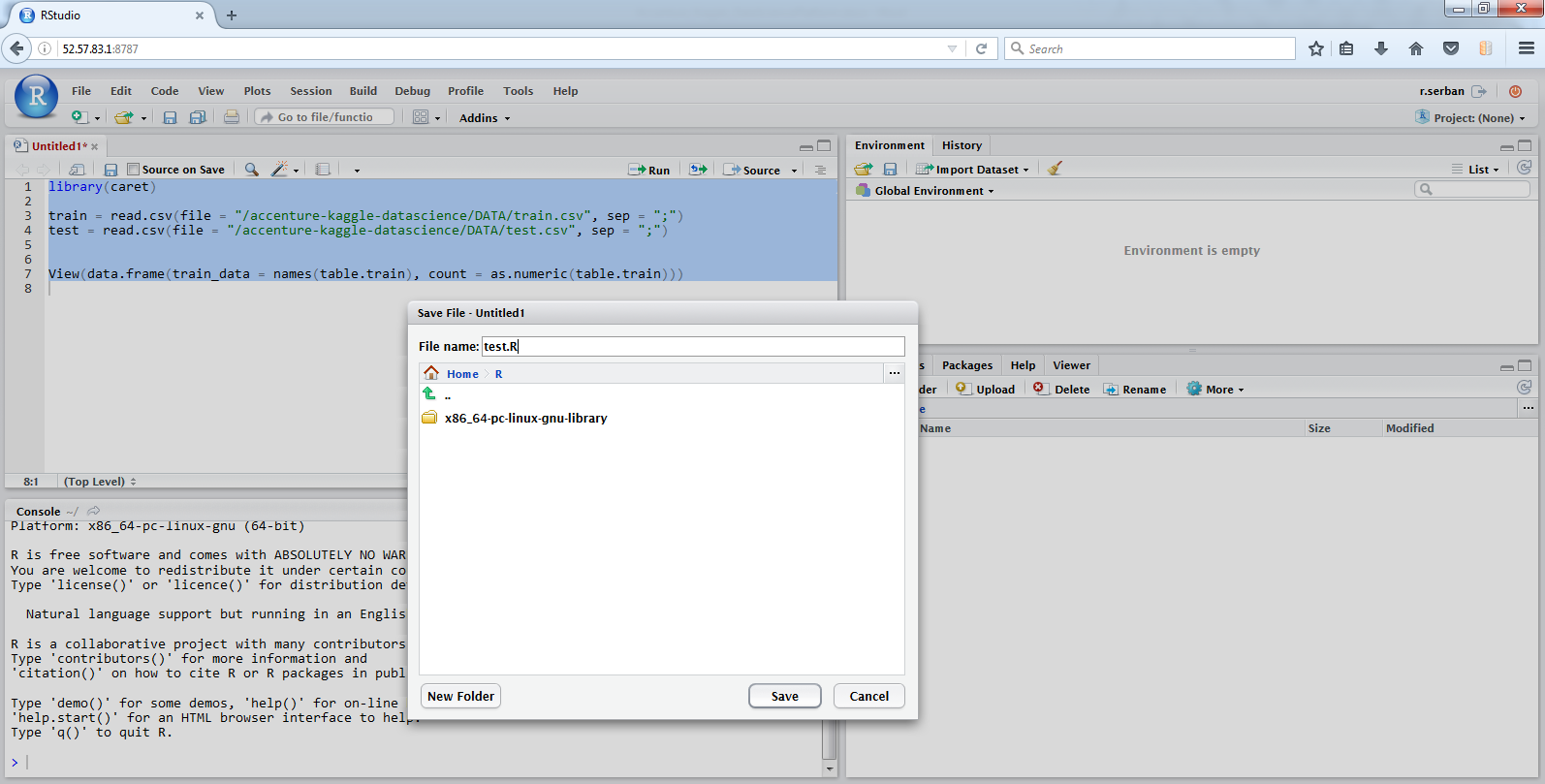
* + You can work in R Studio server (almost) as if you are running local RStudio. Note: please install RStudio (<https://www.rstudio.com/products/rstudio/download/> ) locally, so that you can work offline.



Default location of your locally installed R libraries will be: /home/<username>/R/x86\_64-pc-linux-gnu-library/3.0. If you want to share it with others, you need to send to [r.serban@accenture.com](mailto:r.serban@accenture.com) a request for R package/library to be placed in /usr/lib/R/library

.libPaths()

* + - [1] "/home/r.serban/R/x86\_64-pc-linux-gnu-library/3.0"
    - [2] "/usr/local/lib/R/site-library"
    - [3] "/usr/lib/R/site-library"
    - [4] "/usr/lib/R/library"



### Kaggle data location: /accenture-kaggle-datascience

* + Kaggle data is located in: /accenture-kaggle-datascience/DATA

Please note that path are case-sensitive in Linux.

* + You can read and write data in this folder.

Please keep the proposed team structure in it

* + - DATA/
      * train.csv
      * test.csv
      * TEAM1/
        + r.serban – data specific to own tests
        + sihan.ding
        + …
      * TEAM2/
        + niklas.andersson
        + …
        + paulien.koeleman
      * TEAM3/
        + laury.van.bedaf
    - Data present in a folder is visible to all teams and people with sub-folders in the same folder, and should not be visible to other sibling folders

### Spark Server: <http://52.57.83.1:4040>

Spark 2.10 Server URL: <http://52.57.83.1:4040>

Test Spark calculation:

* cd $SPARK\_HOME
* ./bin/run-example SparkPi 10

Test Spark interactively:

* cd $SPARK\_HOME
* ./bin/spark-shell./bin/spark-shell

Test Spark from Hadoop.

|  |  |
| --- | --- |
| 1 | $ bin/spark-shell |
| 1  2  3 | scala> var file = sc.textFile("hdfs://IP:8020/path/to/textfile.txt")  scala> file.flatMap(line => line.split(",")).map(word => (word, 1)).reduceByKey(\_+\_)  scala> count.saveAsTextFile("hdfs://IP:8020/path/to/ouput") | |

### Git repository: /data/opt/git

### KNIME program for data preparation: /data/opt/knime

### JupyterHub server: traditional IPython style data science notebooks <http://52.57.83.1:8000/>

### Work in progress

* Just in case you have big data: Single-node Hadoop and Spark is also installed on the server