

Snap4city - Quick Guide

List of available tools:

- **Pentaho Kettle**, used to realize ETL processes
- **Node-RED**, used to realize other kind of ETL processes (mainly on data coming from IoT)
- **RStudio**, tool used for statistical analysis

ETL Snap4city - Quick Guide

Introduction

The snap4city ETL process can upload, transform and manage data: such as downloading a file from an external data source, extracting its contents and save in a database or in the file system, etc.

A set of ETL simple processes has been realized to test some of these aspects:

Type	ETL_name	Data License
HTML	Florence_firstAid_accesses_HTML	Open data
Json	Florence_Parking_JSON (static & realTime)	Open data
GeoJson	via_francigena_farmhouse_GeoJson	Under Regione Toscana authorization
XML	Florence_Weather_XML	Open data
Csv	Florence_Pharmacies_CSV	Open data
XLS	Helsinki_youth_subsidies_XLS	Open data
Kmz	Electric_vehicle_charging_kmz	Open data
Shape	Bike_Sharing_Areas_Shp	Open data
gtfs	Tpl_bus_gtfs	Open data
linkedData	Smartbench_LinkedData	Under Comune di Firenze authorization

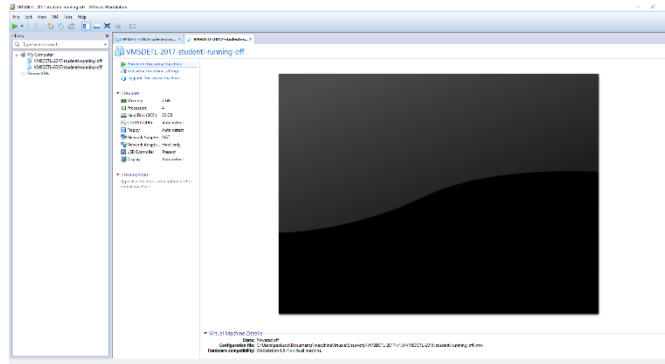
Type	ETL_name	License
HTTP	Florence_firstAid_accesses_HTML	Open data
FTP	Florence_School_canteen	Open data
Datex II	Tuscany_parking	Under MIIC autothorization
Rest API	via_francigena_farmhouse_GeoJson	Under Regione Toscana authorization

JDBC	via_francigena_farmhouse_GeoJson	Under Regione Toscana authorization
SOAP	Tuscany_parking	Under Regione Toscana authorization
M2M	Smartbench_LinkedData	Not public

Virtual machine (VM) access:

The Virtual Machine can be executed with VMware player or workstation, with the following credentials:

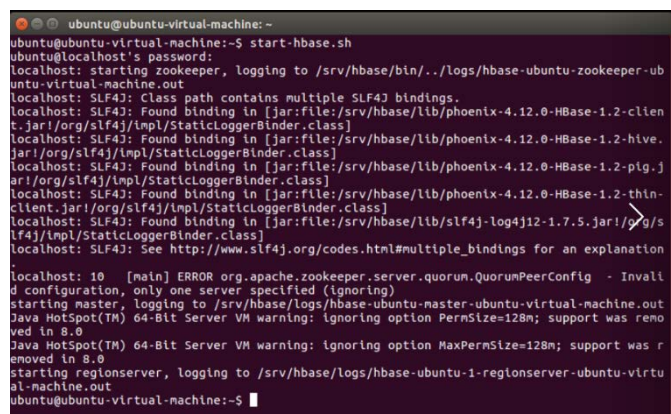
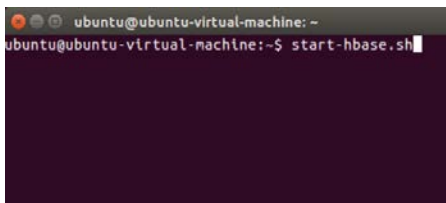
- User: ubuntu
- Password: ubuntu



Tools to use and related commands (from the terminal):

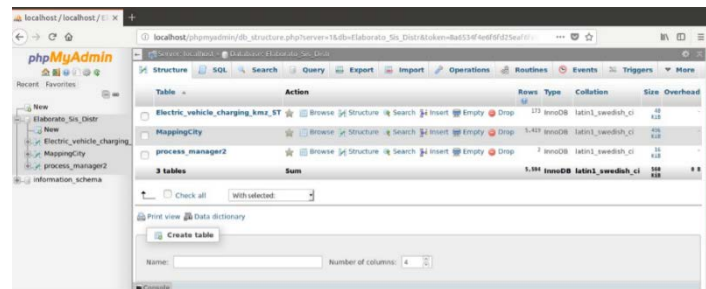
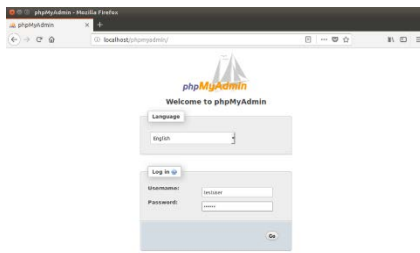
Follow the following steps:

1. Connect to HBase:
 - To run HBase (from any directory): "start-hbase.sh"
 - To stop HBase (from any directory): "stop-hbase.sh"
 - To check the execution: "jps"
 - To check the execution from web interface, visit the web page: <http://localhost:16010/master.jsp>



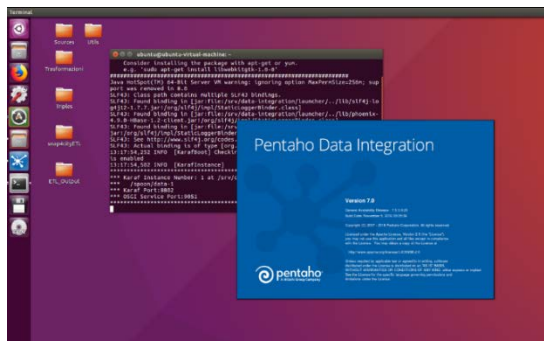
2. Connect to Mysql:

- Use browser: PhpmyAdmin interface <http://127.0.0.1/phpmyadmin/> with credentials:
 - Simple user: username: testuser | password: testpw
 - Root: username: root | password: toor



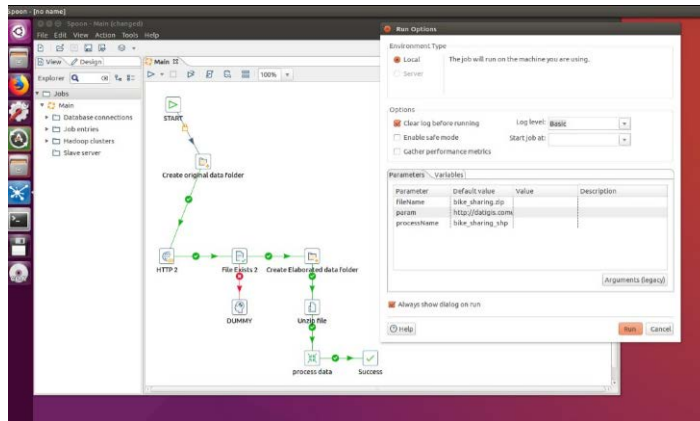
3. Open Spoon:

- To start Spoon, run the following command from any directory on the virtual machine:
 - `spoon.sh`




4. Load an ETL process in spoon

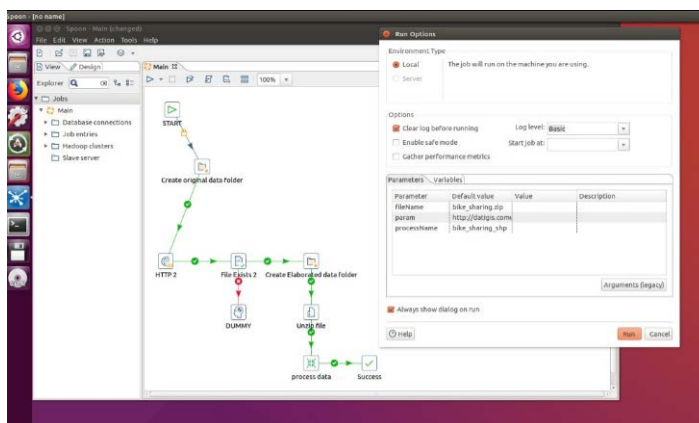
- From the Menu:
 - File> Open > select the file 'Main.kjb' (All the ETL are in the folder: Desktop/snap4cityETL)
 - Example: File > Open > "/home/ubuntu/Desktop/snap4cityETL/Bike_Sharing_Areas_Shp/Main.kjb"



The process is graphically represented as a sequence of elements (or steps), beginning with the 'Start' step and ending with the 'Success' step. Parameters are also associated with the process.

5. Start an ETL process in spoon:

- Click on the Main.jkb 'run' button ()
- A run option window will open with a set of predefined parameters. Then: click on the 'Run' button and the process will start.



6. Work with HBase:

- From a shell, to start HBase: execute the command:
 - 'start-hbase' (as explained above)
- From a shell, to write a table in HBase:
 - Create t1='Francigena_farmhouses','Family1'
- From Spoon, step HBaseInput or HBaseOutput:
 - Put the following params
 - In the tab 'Create mapping', select the right table (e.g. 'Francigena_farmhouses'), click on the button 'Get Incoming fields' and save the mapping.
 - In the tab 'Configure connection', save all

Hadoop cluster

Cluster Name:

Use MapR client

HDFS

Hostname: Port:

Username: Password:

JobTracker

Hostname: Port:

ZooKeeper

Hostname: Port:




Oozie

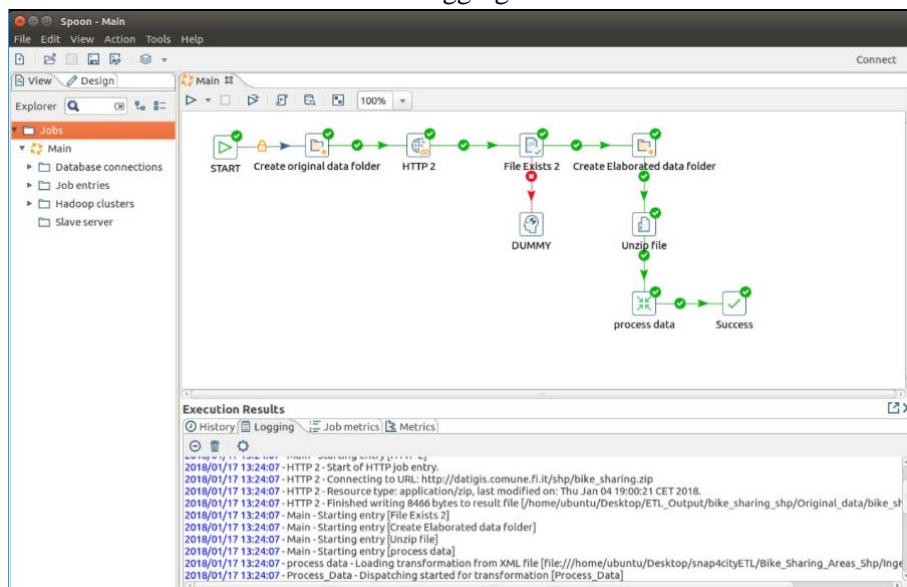
URL:

Results

- ✓ Active Shim Load
Successfully loaded the hdp24 shim.
- ⚠ Shim Configuration Verification
The Hadoop File System URL does not match the URL in the shims core-site.xml.
[Learn more](#)
- ✗ Hadoop File System Connection
Unable to connect to the host.
[Learn more](#)
- ⚠ User Home Directory Access
This test was skipped because Hadoop File System Connection was not successful.
- ⚠ Root Directory Access
This test was skipped because Hadoop File System Connection was not successful.
- ⚠ Verify User Home Permissions
This test was skipped because User Home Directory Access was not successful.
- ✗ Ping Job Tracker / Resource Manager
Unable to connect to the host.
[Learn more](#)
- ✗ Oozie Host Connection
Unable to connect to the host.
[Learn more](#)
- ✓ Zookeeper Ensemble Connection
Connected to all Zookeeper nodes.

7. Validate the success of an ETL:
 In order to validate the success you can:

- 1) See it from the Spoon Editor:
 - Each step has a check symbol in the upper right corner ( ). The symbol is green if the step has been executed in the correct modality, red if something goes wrong. The ETL ends correctly if the final step the 'Success' ends correctly. 
 - In the editor it is also present a window in which the logs are reported:
 - 'Execution results' > Logging

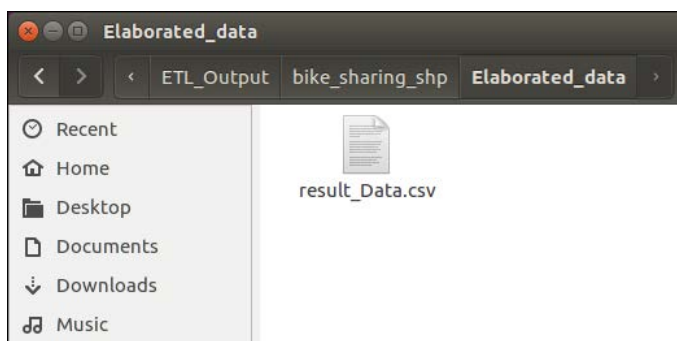
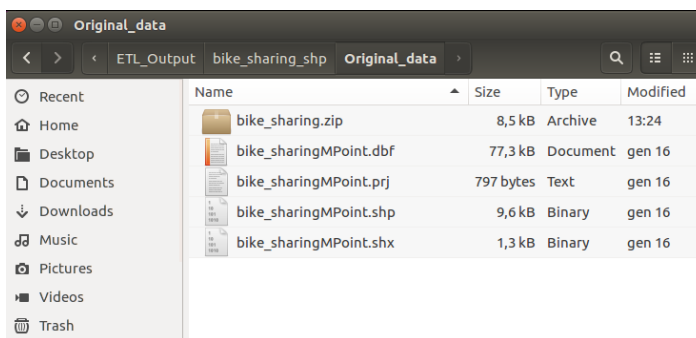
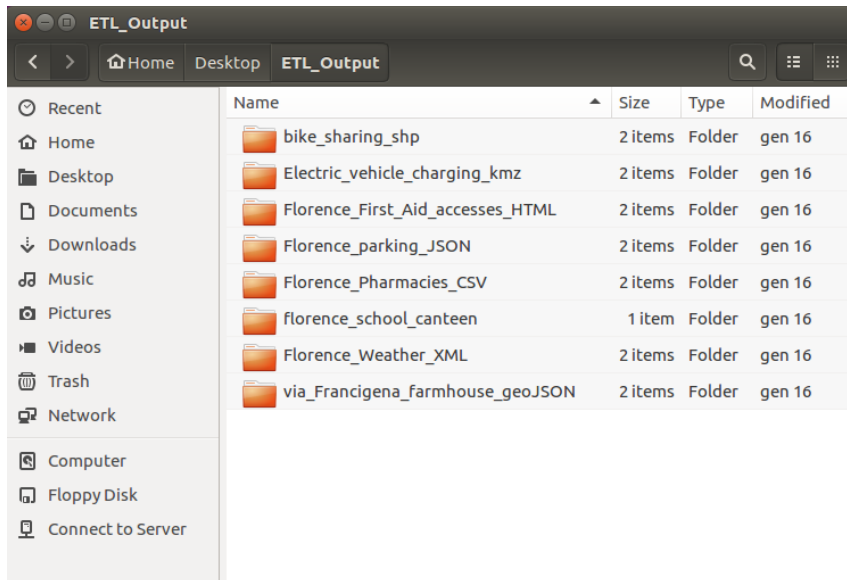


The screenshot shows the Spoon Editor interface. The top part displays a job design with several steps: START, Create original data folder, HTTP 2, File Exists 2, Create Elaborated data folder, DUMMY, Unzip file, process data, and Success. Each step has a green checkmark in its top right corner, indicating successful execution. The bottom part of the screenshot shows the 'Execution Results' window, which is currently displaying the 'Logging' tab. The log contains the following entries:

```

2018/01/17 13:24:07 - HTTP 2 - Start of HTTP job entry.
2018/01/17 13:24:07 - HTTP 2 - Connecting to URL: http://datigis.comune.fi.it/shp/bike_sharing.zip
2018/01/17 13:24:07 - HTTP 2 - Resource type: application/zip, last modified on: Thu Jan 04 19:00:21 CET 2018.
2018/01/17 13:24:07 - HTTP 2 - Finished writing 6466 bytes to result file [/home/ubuntu/Desktop/ETL_Output/bike_sharing_shp/Original_data/bike_sh
2018/01/17 13:24:07 - Main - Starting entry [File Exists 2]
2018/01/17 13:24:07 - Main - Starting entry [Create Elaborated data folder]
2018/01/17 13:24:07 - Main - Starting entry [Unzip file]
2018/01/17 13:24:07 - Main - Starting entry [process data]
2018/01/17 13:24:07 - process data - Loading transformation from XML file [file:///home/ubuntu/Desktop/snapcity/ETL/Bike_Sharing_Areas_Shp/Inge
2018/01/17 13:24:07 - Process_Data - Dispatching started for transformation [Process_Data]
  
```

- 2) See it from the Output folder:
 - The snap4city ETLs are created to demonstrate the different kind of data / protocols used to download, transform and manage data. So, each of them will download a dataset and will put it in the file system. All the files will be contained in the directory: /Desktop/ETL_Output

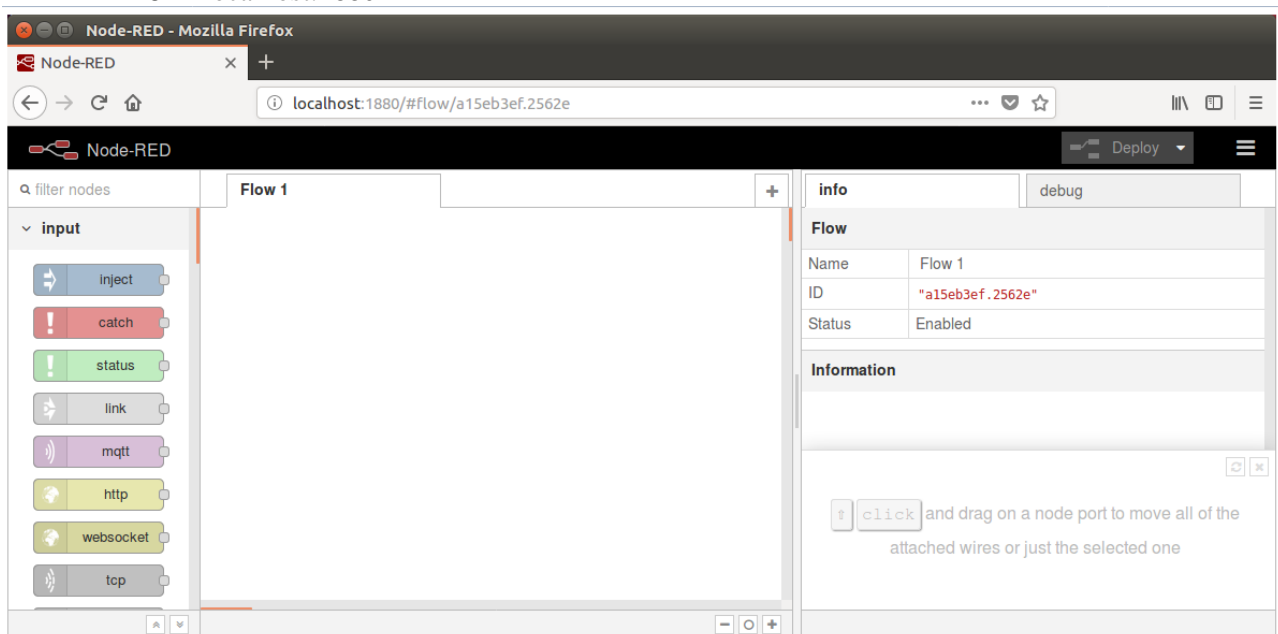


The extended ETL user manual is available here: <http://www.disit.org/7115>

Node-RED - Quick Guide

In the VM it is possible to use the Node-RED tool. In order to do this, it is necessary to launch the tool from the terminal, using the following commands:

- From the /home:
 - Start Node-RED: `./start-nodered.sh`
 - Stop Node-RED: `./stop-nodered.sh`
- The service is available in the browser:
 - Localhost:1880



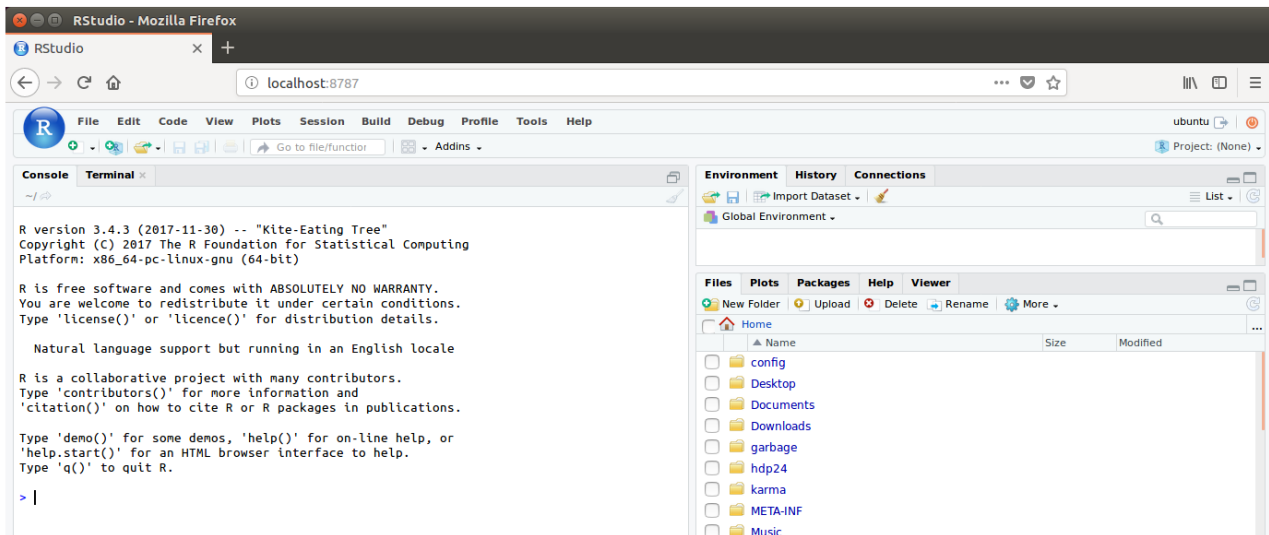
The extended Node-RED user manual is available here: <http://www.disit.org/7112>

RStudio - Quick Guide

In the VM it is possible to use the RStudio

<http://localhost:8787>

- **user: ubuntu**
- **pwd: ubuntu**



The extended RStudio description is available here: <https://www.rstudio.com>