**Web scrapers set up instructions:**

**Web-scrapers:**

The web scrapers code, the run and set up instructions can be found within the bit bucket accounts:

Username: [onswebscraping@gmail.com](mailto:onswebscraping@gmail.com)

Password: ScrapeMyWeb

<https://bitbucket.org/rwb0000/supermarket_scraper>.

Username: [onsbigdata@gmail.com](mailto:onsbigdata@gmail.com)

Password: nibelung14

And the gmail drive on the account:

Username: [onswebscraping@gmail.com](mailto:onswebscraping@gmail.com)

Password: ScrapeMyWeb

* Git (source code management) client to access code in remote repository (BitBucket) e.g.

sudo apt-get install git

**Run system:**

This run system has a number of requirements. Below are instructions on how to set up the environment so that the system will run.

Please do not update Firefox.

## **Software Required:**

The following are needed to run the basic web-scraping process (on Linux):

* Linux Mint (recommended)
* Python 2.7.x (should already be installed on any modern Linux system)
* Python PIP package installer e.g. on Ubuntu/Debian run:
  + sudo apt-get install python-pip
* Install Anaconda
  + https://www.continuum.io/downloads
* Install datetime
  + conda install datetime
* Install pandas
  + conda install pandas
* Download mongoDB
  + sudo apt-get install mongodb
* Install R for linux: <http://cran.us.r-project.org/> version (>3.1). Library plyr will not work on an older versions.

Install R packeges: sudo install R

* To update your R:
* sudo apt-get update
* sudo apt-get install r-base

Install the package plyr. Open R (Sudo R) and type

* install.packages(“plyr”)

If you have an older version of R remove it completely, update and then download an up to-date version:

<http://stackoverflow.com/questions/16093331/how-to-install-r-version-3>

**Database:**

The Mongo database is running on BigData mongo instant. If you cannot connect with the Mongo database then check that the instant is still running. Also check the permissions of the file writing from the folder where the data is (if the permissions are not correctly set then it will still work if you run it through the terminal, but will not run through Crontab).

To change the permissions use the following commands for python and bash files:

chmod +x /...add path here.../\*.sh

chmod +x /...add path here.../\*.py

**Paths:**

A number of paths across the files need to be changed in order for the system to work. The paths will need to be changed to the new location of the files. These files are:

* WS2system\_launcher.py
* WS2system\_launcher2.py
* functions.first\_mongo\_pull\_v3
* functions.anomaly\_detection\_v3
* identify\_missing\_v3.R
* impute\_pack\_stack\_rack\_v3.py
* averaging\_v3.py
* All price index code

**Crontab:**

Use crontab to automate the process.

To change/see the crontab open a terminal from the run\_system folder. Open crontab (type crontab -e to edit). The current automated run files will be present. These are the commands that run the web scrapers so please do not turns these off/change them unless you are sure of the results. The run command looks like this:

* 37 10 \* \* \* python /media/mint/e834712c-23da-4cbe-a4ec-3a35d416877b/Run\_system/WebScrapingSystem/WS2system\_launcher.py

The first part of this command 37 10 \* \* \* means: min hour day month year. Anything set to \* will run every one of these time periods. So 37 10 \* \* \* will run every day at 10.37am.

To add the crontab to automate your system, open crontab, add two paths. One two the first path of the system launcher, and then another to the second. Leave at least 20 minutes between the two so that the first has time to finish running. Also, make sure this is run after the web scrapers have finished so that the most up-to-date data is used.

E.g.

* 30 10 \* \* \* python /..add your path here/WS2system\_launcher.py
* 00 11 \* \* \* python /..add your path here/WS2system\_launcher2.py