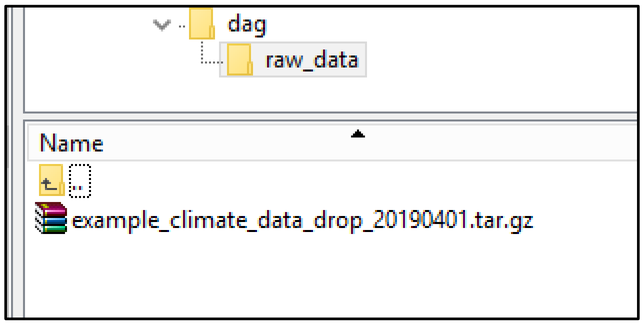
DA Global CONSULTANTS

Dynamically Import Flat Files Using Your OS and Python

By harnessing the power of your Linux OS, you can run code in your Python scripts to manipulate your flat files and import them into your Analytics Environment.

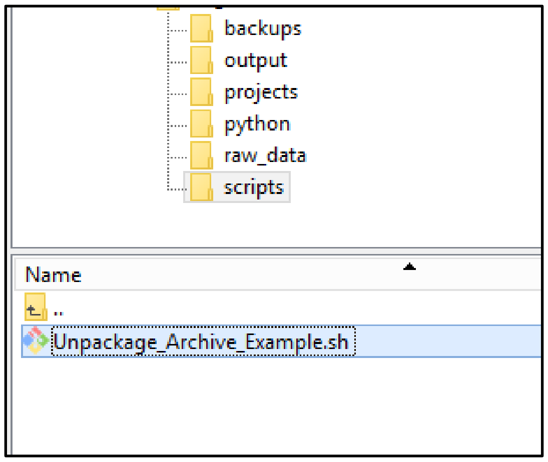
We will start off with a package of files that we have received and have dropped into our file system on our Linux OS.

example_climate_data_drop_20190401.tar.gz



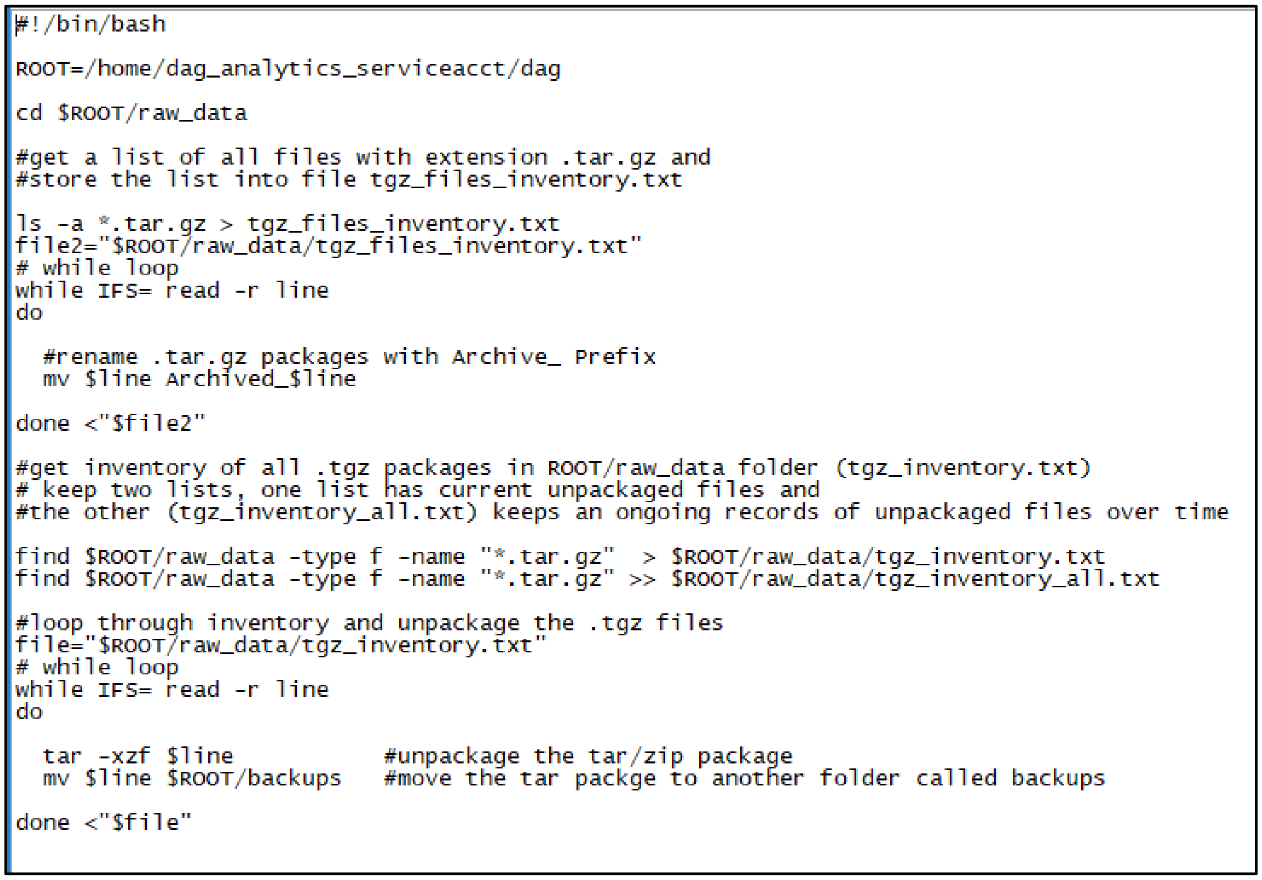
First we will build a script that we can execute in the OS to un-package the files and move/archive the tar package for backups in case we need to restore the data or for future use.

Unpackage_Archive_Example.sh will be stored in the scripts root folder



This Script will:

- Rename the tar.gz archive file with prefix **Archive_**<tar.gz file name>.
- Store the file name(s) into a txt file to reference.
- Un-package the contents of the tar.gz file(s).
- Move the tar.gz package to a backup location.



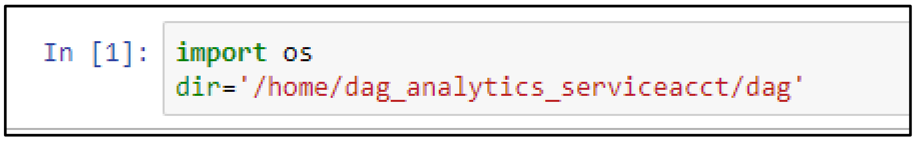
```bash
#!/bin/bash

ROOT=/home/dag_analytics_serviceacct/dag

cd $ROOT/raw_data

#get a list of all files with extension .tar.gz and
#store the list into file tgz_files_inventory.txt

ls -a *.tar.gz > tgz_files_inventory.txt
file2="$ROOT/raw_data/tgz_files_inventory.txt"
# while loop
while IFS= read -r line
do

   #rename .tar.gz packages with Archive_ Prefix
   mv $line Archived_$line

done <"$file2"

#get inventory of all .tgz packages in ROOT/raw_data folder (tgz_inventory.txt)
# keep two lists, one list has current unpackaged files and
#the other (tgz_inventory_all.txt) keeps an ongoing records of unpackaged files over time

find $ROOT/raw_data -type f -name "*.tar.gz"  > $ROOT/raw_data/tgz_inventory.txt
find $ROOT/raw_data -type f -name "*.tar.gz" >> $ROOT/raw_data/tgz_inventory_all.txt

#loop through inventory and unpackage the .tgz files
file="$ROOT/raw_data/tgz_inventory.txt"
# while loop
while IFS= read -r line
do

   tar -xzf $line           #unpackage the tar/zip package
   mv $line $ROOT/backups   #move the tar packge to another folder called backups

done <"$file"
```

Now let us run this script from Python.

We will use the Jupyter Notebook to run our python commands.

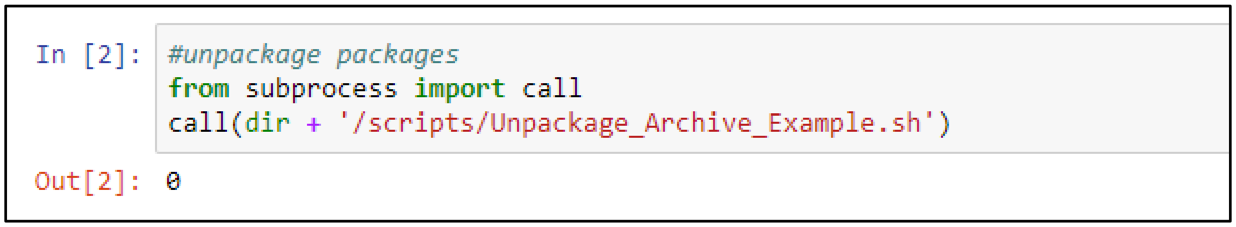
The First Code Snippet will run 2 components:

1. Import our Operating System Module (os) so that we can run commands on our Linux environment via our python code.
2. Set a variable named dir, which represents the root location of our project for future use. This is especially handy if we had to port our project to a different file location to minimize code changes.



```
In [1]: import os
        dir='/home/dag_analytics_serviceacct/dag'
```

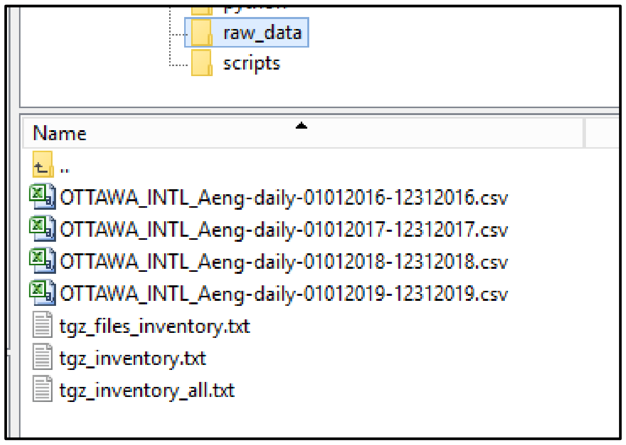
This next cell will call the script we created Unpackage_Archive_Example.sh and run it.



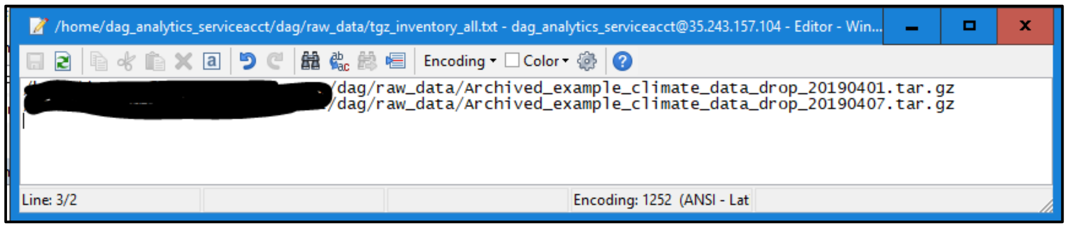
```
In [2]:  #unpackage packages
         from subprocess import call
         call(dir + '/scripts/Unpackage_Archive_Example.sh')

Out[2]:  0
```

The results from running our script:

1. Unpackaged the example_climate_data_drop_20190401.tar.gz which contained multiple **.csv** files
2. Created .txt output files which have recorded what tar.gz file has been unpackaged to keep a record for future use. In the file **tgz_inventory_all.txt** will list all the tar.gz files that have ever been unpackaged by this process. Below there is an example of the output in tgz_inventory_all.txt which indicates that we have unpackaged 2 packages over time.
3. Original tar.gz packages have been renamed with a prefix Archived_ and moved into our backups folder.



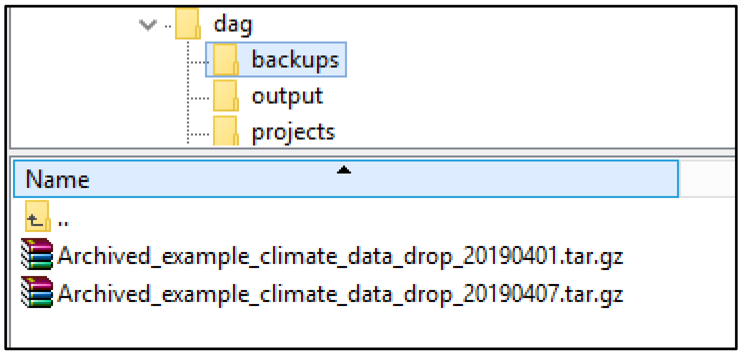
```
raw_data
scripts

Name
..
OTTAWA_INTL_Aeng-daily-01012016-12312016.csv
OTTAWA_INTL_Aeng-daily-01012017-12312017.csv
OTTAWA_INTL_Aeng-daily-01012018-12312018.csv
OTTAWA_INTL_Aeng-daily-01012019-12312019.csv
tgz_files_inventory.txt
tgz_inventory.txt
tgz_inventory_all.txt
```



/home/dag_analytics_serviceacct/dag/raw_data/tgz_inventory_all.txt - dag_analytics_serviceacct@35.243.157.104 - Editor - Win...

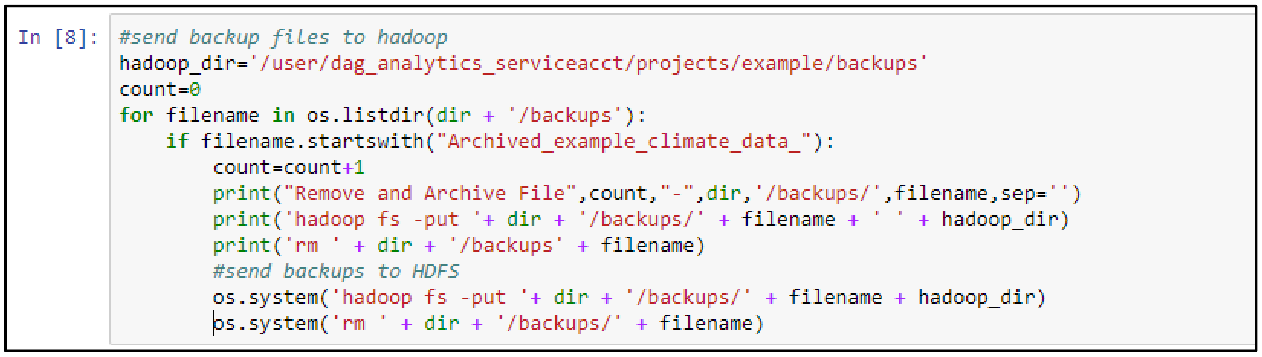
```
[redacted]/dag/raw_data/Archived_example_climate_data_drop_20190401.tar.gz
/dag/raw_data/Archived_example_climate_data_drop_20190407.tar.gz
```

Line: 3/2   Encoding: 1252 (ANSI - Lat



```
Archived_example_climate_data_drop_20190401.tar.gz
Archived_example_climate_data_drop_20190407.tar.gz
```

The Last Snippet in this project will move the archived tar.gz file(s) from its location in the backups directory on the OS file system and move it into our Hadoop HDFS for permanent storage.



```
In [8]: #send backup files to hadoop
        hadoop_dir='/user/dag_analytics_serviceacct/projects/example/backups'
        count=0
        for filename in os.listdir(dir + '/backups'):
            if filename.startswith("Archived_example_climate_data_"):
                count=count+1
                print("Remove and Archive File",count,"-",dir,'/backups/',filename,sep='')
                print('hadoop fs -put '+ dir + '/backups/' + filename + ' ' + hadoop_dir)
                print('rm ' + dir + '/backups' + filename)
                #send backups to HDFS
                os.system('hadoop fs -put '+ dir + '/backups/' + filename + hadoop_dir)
                os.system('rm ' + dir + '/backups/' + filename)
```

I have embedded print statements into this process so that we can verify our locations.

The Print Output commands results from the above snippet for moving our 2 example archived backup files into the HDFS and removing them the backups directory.

```
 Remove and Archive File1-
/home/dag_analytics_serviceacct/dag/backups/Archived_example_climate_data_drop_20190401.tar.gz

hadoop fs -put
/home/dag_analytics_serviceacct/dag/backups/Archived_example_climate_data_drop_20190401.tar.gz
/user/dag_analytics_serviceacct/projects/example/backups

rm /home/dag_analytics_serviceacct/dag/backupsArchived_example_climate_data_drop_20190401.tar.gz

Remove and Archive File2-
/home/dag_analytics_serviceacct/dag/backups/Archived_example_climate_data_drop_20190407.tar.gz

hadoop fs -put
/home/dag_analytics_serviceacct/dag/backups/Archived_example_climate_data_drop_20190407.tar.gz
/user/dag_analytics_serviceacct/projects/example/backups

rm /home/dag_analytics_serviceacct/dag/backupsArchived_example_climate_data_drop_20190407.tar.gz
```