python_hadoop_pig

July 20, 2022

1 Python Hadoop Pig

This notebook aims at showing how to submit a PIG job to remote hadoop cluster (tested with Cloudera). It works better if you know Hadoop otherwise I recommend reading Map/Reduce avec PIG (French). First, we download data. We are going to upload that data to the remote cluster. The Hadoop distribution tested here is Cloudera.

[1]:
```
import pyensae
%load_ext pyensae
%load_ext pyenbc
pyensae.download_data("ConfLongDemo_JSI.txt", website="https://archive.ics.uci.edu/ml/
  ↪machine-learning-databases/00196/")
```

[1]:
```
'ConfLongDemo_JSI.txt'
```

We open a SSH connection to the bridge which can communicate to the cluster.

[2]:
```
import pyquickhelper.ipythonhelper as ipy
params={"server":"", "username":"", "password":""}
ipy.open_html_form(params=params,title="credentials",key_save="ssh_remote_hadoop")
```

[2]:
```
<IPython.core.display.HTML at 0x742c9f0>
```

[3]:
```
password = ssh_remote_hadoop["password"]
server = ssh_remote_hadoop["server"]
username = ssh_remote_hadoop["username"]
```

We open the SSH connection:

[4]:
```
%remote_open
```

[4]:
```
<pyensae.remote.ssh_remote_connection.ASSHClient at 0xa2422e8>
```

We check the content of the remote machine:

[5]:
```
%remote_cmd ls -l
```

[5]:
```
<IPython.core.display.HTML object>
```

[6]:
```
%remote_ls .
```

[6]:
```
               attributes         code    alias folder  size   unit  \
-rw-rw-r-- 1  xavierdupre  xavierdupre     1043    Jul    14  23:40
-rw-r--r-- 1  xavierdupre  xavierdupre        2    Jul    15  00:22
```

```
-rw-rw-r--  1  xavierdupre  xavierdupre        0    Sep    27  00:21
            1  xavierdupre  xavierdupre      290    Jul    14  23:48
            1  xavierdupre  xavierdupre     1654    Jul    15  00:20
            1  xavierdupre  xavierdupre      235    Jul    14  23:37
            1  xavierdupre  xavierdupre     1778    Jul    14  23:57
            1  xavierdupre  xavierdupre     4570    Jul    15  00:45
            1  xavierdupre  xavierdupre     4570    Jul    15  23:52
            1  xavierdupre  xavierdupre      574    Jul    15  23:51
            1  xavierdupre  xavierdupre      659    Sep    27  00:21
            1  xavierdupre  xavierdupre      382    Sep    27  00:21
            1  xavierdupre  xavierdupre    26186    Jul    15  23:52
            1  xavierdupre  xavierdupre        0    Jul    15  23:51
            1  xavierdupre  xavierdupre  3400818    Jul    15  23:48

                                    name  isdir
-rw-rw-r--  1     centrer_reduire.pig  False
-rw-r--r--  1            diff_cluster  False
-rw-rw-r--  1                   dummy  False
            1         init_random.pig  False
            1  iteration_complete.pig  False
            1      nb_obervations.pig  False
            1   pig_1436911046432.log  False
            1   pig_1436913856496.log  False
            1   pig_1436997076356.log  False
            1     post_traitement.pig  False
            1            pystream.pig  False
            1             pystream.py  False
            1         redirection.err  False
            1         redirection.out  False
            1        Skin_NonSkin.txt  False
```

We check the content on the cluster:

[7]:
```
%remote_cmd hdfs dfs -ls
```

[7]:
```
<IPython.core.display.HTML object>
```

[8]:
```
%dfs_ls .
```

[8]:

	attributes	code	alias	folder	size	date	time	\
0	drwx------	-	xavierdupre	xavierdupre	0	2015-09-27	02:00	
1	drwx------	-	xavierdupre	xavierdupre	0	2015-09-27	00:22	
2	-rw-r--r--	3	xavierdupre	xavierdupre	132727	2014-11-16	02:37	
3	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2014-11-16	02:38	
4	-rw-r--r--	3	xavierdupre	xavierdupre	3400818	2015-07-14	23:35	
5	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-15	00:22	
6	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-14	23:44	
7	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-14	23:43	
8	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-14	23:49	
9	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-14	23:41	
10	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-14	23:38	
11	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-15	00:05	
12	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-15	00:22	
13	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-15	00:07	
14	drwxr-xr-x	-	xavierdupre	xavierdupre	0	2015-07-15	00:09	

```
15  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:11
16  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:13
17  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:15
18  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:17
19  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:18
20  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-07-15  00:20
21  -rw-r--r--    3  xavierdupre  xavierdupre   461444  2014-11-20  01:33
22  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-23  22:03
23  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-23  22:07
24  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-12-03  22:55
25  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-20  23:43
26  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-09-27  00:23
27  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2015-09-27  00:22
28  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-20  01:53
29  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-21  01:17
30  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-23  21:34
31  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-23  21:51
32  drwxr-xr-x    -  xavierdupre  xavierdupre        0  2014-11-21  11:08
```

	name	isdir
0	.Trash	True
1	.staging	True
2	ConfLongDemo_JSI.small.example.txt	False
3	ConfLongDemo_JSI.small.example2.walking.txt	True
4	Skin_NonSkin.txt	False
5	diff_cluster	True
6	donnees_normalisees	True
7	ecartstypes	True
8	init_random	True
9	moyennes	True
10	nb_obervations	True
11	output_iter1	True
12	output_iter10	True
13	output_iter2	True
14	output_iter3	True
15	output_iter4	True
16	output_iter5	True
17	output_iter6	True
18	output_iter7	True
19	output_iter8	True
20	output_iter9	True
21	paris.2014-11-11_22-00-18.331391.txt	False
22	python_info.txt	True
23	python_info2.txt	True
24	random	True
25	unitest2	True
26	unittest	True
27	unittest2	True
28	velib_1hjs	True
29	velib_py	True
30	velib_py_results	True
31	velib_py_results_3days	True
32	velib_several_days	True

We upload the file on the bridge (we should zip it first, it would reduce the uploading time).

[9]:
```
%remote_up ConfLongDemo_JSI.txt ConfLongDemo_JSI.txt
```

[9]:
```
'ConfLongDemo_JSI.txt'
```

We check it got there:

[10]:
```
%remote_cmd ls Conf*JSI.txt
```

[10]:
```
<IPython.core.display.HTML object>
```

We put it on the cluster:

[11]:
```
%remote_cmd hdfs dfs -put ConfLongDemo_JSI.txt ConfLongDemo_JSI.txt
```

[11]:
```
<IPython.core.display.HTML object>
```

We check it was put on the cluster:

[12]:
```
%remote_cmd hdfs dfs -ls Conf*JSI.txt
```

[12]:
```
<IPython.core.display.HTML object>
```

[13]:
```
dfs_ls Conf*JSI.txt
```

[13]:
```
   attributes code        alias       folder      size        date   time  \
0  -rw-r--r--    3  xavierdupre  xavierdupre  21546346  2015-09-27  11:33

                   name  isdir
0  ConfLongDemo_JSI.txt  False
```

We create a simple PIG program:

[14]:
```
%%PIG filter_example.pig

myinput = LOAD 'ConfLongDemo_JSI.txt' USING PigStorage(',') AS
    (index:long, sequence, tag, timestamp:long, dateformat, x:double,y:double, z:
 ↪double, activity) ;
filt = FILTER myinput BY activity == 'walking' ;
STORE filt INTO 'ConfLongDemo_JSI.walking.txt' USING PigStorage() ;
```

[15]:
```
%pig_submit filter_example.pig -r=filter_example.redirect
```

[15]:
```
<IPython.core.display.HTML object>
```

We check the redirected files were created:

[16]:
```
%remote_cmd ls f*redirect*
```

[16]:
```
<IPython.core.display.HTML object>
```

We check the tail on a regular basis to see the job running (some other commands can be used to monitor jobs, `%remote_cmd mapred --help`).

[17]:
```
%remote_cmd tail filter_example.redirect.err
```

[17]:
```
<IPython.core.display.HTML object>
```

[18]:
```
%remote_cmd hdfs dfs -ls Conf*JSI.walking.txt
```

[18]:
```
<IPython.core.display.HTML object>
```

[19]:
```
%dfs_ls Conf*JSI.walking.txt
```

[19]:
```
   attributes code        alias       folder size        date   time  \
0  -rw-r--r--    3  xavierdupre  xavierdupre    0  2015-09-27  11:38
1  -rw-r--r--    3  xavierdupre  xavierdupre    0  2015-09-27  11:38

                                       name  isdir
0      ConfLongDemo_JSI.walking.txt/_SUCCESS  False
1  ConfLongDemo_JSI.walking.txt/part-m-00000  False
```

After that, the stream has to downloaded to the bridge and then to the local machine with `%remote_down`. We finally close the connection.

[20]:
```
%remote_close
```

[20]:
```
True
```

END

[21]:
```

```