pig_azure_correction

July 1, 2023

1 Map/Reduce avec PIG sur Azure - correction

[1]:
```
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[1]:
```
<IPython.core.display.HTML object>
```

1.1 Données

On considère le jeu de données suivant : Localization Data for Person Activity Data Set qu'on récupère comme indiqué dans le notebook de l'énoncé.

[2]:
```
from pyquickhelper.ipythonhelper import open_html_form
params={"blob_storage":"", "password1":"", "hadoop_server":"", "password2":"",␣
↪"username":"xavierdupre"}
open_html_form(params=params,title="server + hadoop + credentials", key_save="blobhp")
```

[2]:
```
<IPython.core.display.HTML object>
```

[3]:
```
blobstorage = blobhp["blob_storage"]
blobpassword = blobhp["password1"]
hadoop_server = blobhp["hadoop_server"]
hadoop_password = blobhp["password2"]
username = blobhp["username"]
```

[4]:
```
import pyensae
%load_ext pyensae
%load_ext pyenbc
%hd_open
```

[4]:
```
(<pyensae.remote.azure_connection.AzureClient at 0xafe7e10>,
 <azure.storage.blob.blobservice.BlobService at 0xafe7e48>)
```

1.2 Exercice 1 : GROUP BY

[5]:
```
import pandas, sqlite3
con = sqlite3.connect("ConfLongDemo_JSI.db3")
df = pandas.read_sql("""SELECT activity, count(*) as nb FROM person GROUP BY␣
↪activity""", con)
con.close()
```

```
df.head()
```

[5]:

```
       activity     nb
0       falling   2973
1         lying  54480
2    lying down   6168
3  on all fours   5210
4       sitting  27244
```

On vérifie que le fichier qu'on veut traiter est bien là :

[6]:

```
%blob_ls /testensae/ConfLongDemo_JSI.small.txt
```

[6]:

```
                                  name                  last_modified  \
0  testensae/ConfLongDemo_JSI.small.txt  Thu, 29 Oct 2015 00:23:00 GMT

                content_type  content_length  blob_type
0  application/octet-stream          132727  BlockBlob
```

Il faut maintenant le faire avec PIG.

[7]:

```
%%PIG_azure solution_groupby.pig

myinput = LOAD '$CONTAINER/testensae/ConfLongDemo_JSI.small.txt'
          using PigStorage(',')
          AS (index:long, sequence, tag, timestamp:long, dateformat, x:double,y:
 ↪double, z:double, activity) ;

gr = GROUP myinput BY activity ;
avgact = FOREACH gr GENERATE group, COUNT(myinput) ;

STORE avgact INTO '$CONTAINER/$PSEUDO/testensae/ConfLongDemo_JSI.small.group.2015.txt'␣
 ↪USING PigStorage() ;
```

On soumet le job :

[8]:

```
jid = %hd_pig_submit solution_groupby.pig
jid
```

[8]:

```
{'id': 'job_1445989166328_0009'}
```

On vérifie le status du job :

[9]:

```
st = %hd_job_status jid["id"]
st["id"],st["percentComplete"],st["completed"],st["status"]["jobComplete"],st["status"]["state"]
```

[9]:

```
('job_1445989166328_0009', '100% complete', None, False, 'RUNNING')
```

On regarde si la compilation s'est bien passée :

[10]:

```
%hd_tail_stderr jid["id"]
```

[10]:

```
<IPython.core.display.HTML object>
```

On regarde le contenu du répertoire sur le blob storage :

[11]:

```
df=%blob_ls /$PSEUDO/testensae
list(df["name"])
```

[11]:
```
['xavierdupre/testensae',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt/part-m-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt/part-m-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt/part-m-00000']
```

[12]:
```
import os
if os.path.exists("results.group.2015.xt") : os.remove("results.group.2015.txt")
%blob_downmerge /$PSEUDO/testensae/ConfLongDemo_JSI.small.group.2015.txt results.group.
 ↪2015.txt
```

[12]:
```
'results.group.2015.txt'
```

[13]:
```
%lsr res.*[.]txt
```

[13]:
```
   directory              last_modified                      name      size
0      False 2015-10-29 01:56:11.025867  .\results.group.2015.txt        89
1      False 2015-10-29 01:46:45.425028             .\results.txt  21.65 Kb
2      False 2015-10-29 01:46:46.705466    .\results_allfiles.txt  21.65 Kb
```

[14]:
```
%head results.group.2015.txt
```

[14]:
```
<IPython.core.display.HTML object>
```

1.3 Exercice 2 : JOIN

[15]:
```
con = sqlite3.connect("ConfLongDemo_JSI.db3")
df = pandas.read_sql("""SELECT person.*, A.nb FROM person INNER JOIN (
                            SELECT activity, count(*) as nb FROM person GROUP BY␣
 ↪activity) AS A
                            ON person.activity == A.activity""", con)
con.close()
df.head()
```

[15]:
```
   index sequence              tag           timestamp  \
0      0      A01  010-000-024-033  633790226051280329
1      1      A01  020-000-033-111  633790226051820913
2      2      A01  020-000-032-221  633790226052091205
3      3      A01  010-000-024-033  633790226052361498
```

```
4      4      A01  010-000-030-096  633790226052631792

                dateformat         x         y         z activity     nb
0  27.05.2009 14:03:25:127  4.062931  1.892434  0.507425  walking  32710
1  27.05.2009 14:03:25:183  4.291954  1.781140  1.344495  walking  32710
2  27.05.2009 14:03:25:210  4.359101  1.826456  0.968821  walking  32710
3  27.05.2009 14:03:25:237  4.087835  1.879999  0.466983  walking  32710
4  27.05.2009 14:03:25:263  4.324462  2.072460  0.488065  walking  32710
```

Idem, maintenant il faut le faire avec PIG.

[16]:
```
%%PIG_azure solution_groupby_join.pig

myinput = LOAD '$CONTAINER/testensae/ConfLongDemo_JSI.small.txt'
          using PigStorage(',')
          AS (index:long, sequence, tag, timestamp:long, dateformat, x:double,y:
 ↪double, z:double, activity) ;

gr = GROUP myinput BY activity ;
avgact = FOREACH gr GENERATE group, COUNT(myinput) ;

joined = JOIN myinput BY activity, avgact BY group ;

STORE joined INTO '$CONTAINER/$PSEUDO/testensae/ConfLongDemo_JSI.small.group.join.2015.
 ↪txt' USING PigStorage() ;
```

[17]:
```
jid = %hd_pig_submit solution_groupby_join.pig
jid
```

[17]:
```
{'id': 'job_1445989166328_0011'}
```

[18]:
```
st = %hd_job_status jid["id"]
st["id"],st["percentComplete"],st["completed"],st["status"]["jobComplete"],st["status"]["state"],␣
 ↪st["userargs"]["file"]
```

[18]:
```
('job_1445989166328_0011',
 '100% complete',
 'done',
 True,
 'SUCCEEDED',
 'wasb://hdblobstorage@hdblobstorage.blob.core.windows.net/xavierdupre/scripts/p
ig/solution_groupby_join.pig')
```

[19]:
```
df=%blob_ls /$PSEUDO/testensae
df
```

[19]:
```
                                                name  \
0                                 xavierdupre/testensae
1   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
2   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
3   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
4   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
5   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
6   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
```

```
7   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
8   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
9   xavierdupre/testensae/ConfLongDemo_JSI.small.g…
10  xavierdupre/testensae/ConfLongDemo_JSI.small.g…
11  xavierdupre/testensae/ConfLongDemo_JSI.small.g…
12  xavierdupre/testensae/ConfLongDemo_JSI.small.g…
13  xavierdupre/testensae/ConfLongDemo_JSI.small.k…
14  xavierdupre/testensae/ConfLongDemo_JSI.small.k…
15  xavierdupre/testensae/ConfLongDemo_JSI.small.k…
16  xavierdupre/testensae/ConfLongDemo_JSI.small.w…
17  xavierdupre/testensae/ConfLongDemo_JSI.small.w…
18  xavierdupre/testensae/ConfLongDemo_JSI.small.w…
19  xavierdupre/testensae/ConfLongDemo_JSI.small.w…
20  xavierdupre/testensae/ConfLongDemo_JSI.small.w…
21  xavierdupre/testensae/ConfLongDemo_JSI.small.w…

                   last_modified              content_type  content_length  \
0   Tue, 25 Nov 2014 00:50:34 GMT  application/octet-stream               0
1   Thu, 29 Oct 2015 00:55:09 GMT                                          0
2   Thu, 29 Oct 2015 00:55:09 GMT  application/octet-stream               0
3   Thu, 29 Oct 2015 00:55:08 GMT  application/octet-stream              89
4   Thu, 29 Oct 2015 00:58:43 GMT                                          0
5   Thu, 29 Oct 2015 00:58:43 GMT  application/octet-stream               0
6   Thu, 29 Oct 2015 00:58:42 GMT  application/octet-stream          144059
7   Tue, 25 Nov 2014 01:16:11 GMT                                          0
8   Tue, 25 Nov 2014 01:16:11 GMT  application/octet-stream               0
9   Tue, 25 Nov 2014 01:16:10 GMT  application/octet-stream          144059
10  Tue, 25 Nov 2014 01:12:49 GMT                                          0
11  Tue, 25 Nov 2014 01:12:49 GMT  application/octet-stream               0
12  Tue, 25 Nov 2014 01:12:49 GMT  application/octet-stream              89
13  Tue, 25 Nov 2014 00:50:45 GMT                                          0
14  Tue, 25 Nov 2014 00:50:46 GMT  application/octet-stream               0
15  Tue, 25 Nov 2014 00:50:45 GMT  application/octet-stream           22166
16  Thu, 29 Oct 2015 00:28:30 GMT                                          0
17  Thu, 29 Oct 2015 00:28:30 GMT  application/octet-stream               0
18  Thu, 29 Oct 2015 00:28:30 GMT  application/octet-stream           22166
19  Thu, 29 Oct 2015 00:46:05 GMT                                          0
20  Thu, 29 Oct 2015 00:46:05 GMT  application/octet-stream               0
21  Thu, 29 Oct 2015 00:46:04 GMT  application/octet-stream           22166

    blob_type
0   BlockBlob
1   BlockBlob
2   BlockBlob
3   BlockBlob
4   BlockBlob
5   BlockBlob
6   BlockBlob
7   BlockBlob
8   BlockBlob
9   BlockBlob
10  BlockBlob
11  BlockBlob
```

```
12  BlockBlob
13  BlockBlob
14  BlockBlob
15  BlockBlob
16  BlockBlob
17  BlockBlob
18  BlockBlob
19  BlockBlob
20  BlockBlob
21  BlockBlob
```

[20]:
```
set(df.name)
```

[20]:
```
{'xavierdupre/testensae',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.2015.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.2015.txt/_SUCCESS',
'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.2015.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.join.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.group.txt/part-r-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.keep_walking.txt/part-m-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking2015.txt/part-m-00000',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt/_SUCCESS',
 'xavierdupre/testensae/ConfLongDemo_JSI.small.walking_2015.txt/part-m-00000'}
```

[21]:
```
if os.path.exists("results.join.2015.txt") : os.remove("results.join.2015.txt")
%blob_downmerge /$PSEUDO/testensae/ConfLongDemo_JSI.small.group.join.2015.txt results.
 ↪join.2015.txt
```

[21]:
```
'results.join.2015.txt'
```

[22]:
```
%head results.join.2015.txt
```

[22]:
```
<IPython.core.display.HTML object>
```

Prolongements

PIG n'est pas la seule façon d'exécuter des jobs Map/Reduce. Hive est un langage dont la syntaxe est très proche de celle du SQL. L'article Comparing Pig Latin and SQL for Constructing Data Processing Pipelines explicite les différences des deux approches.

langage haut niveau

Ce qu'il faut retenir est que le langage PIG est un langage haut niveau. Le programme est compilé en une séquence d'opérations Map/Reduce transparente pour l'utilisateur. Le temps de développement est très réduit lorsqu'on le compare au même programme écrit en Java. Le compilateur construit un plan d'exécution

(quelques exemples ici) et infère le nombre de machines requises pour distribuer le job. Cela suffit pour la plupart des besoins, cela nécessite.

petits jeux

Certains jobs peuvent durer des heures, il est conseillée de les essayer sur des petits jeux de données avant de les faire tourner sur les vrais données. Il est toujours frustrant de s'apercevoir qu'un job a planté au bout de deux heures car une chaîne de caractères est vide et que ce cas n'a pas été prévu.

Avec ces petits jeux, il est possible de faire tourner et conseillé de tester le job d'abord sur la passerelle (exécution local) avant de le lancer sur le cluster. Avec pyensae, il faut ajouter l'option `-local` à la commande hd_pig_submit.

concaténer les fichiers divisés

Un programme PIG ne produit pas un fichier mais plusieurs fichiers dans un répertoire. La commande getmerge télécharge ces fichiers sur la passerelle et les fusionne en un seul.

ordre des lignes

Les jobs sont distribués, même en faisant rien (LOAD + STORE), il n'est pas garanti que l'ordre des lignes soit préservé. La probabilié que ce soit le cas est quasi nulle.

[23] :

```
```