

azure_pig

November 26, 2021

1 HDInsight, PIG

Short examples on how to connect to a cluster from a notebook and submit a job (Azure + PIG).

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

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

1.1 Download the data

```
[2]: url = "https://archive.ics.uci.edu/ml/machine-learning-databases/00222/"
      file = "bank.zip"
      import pyensae
      data = pyensae.download_data(file, website=url)
```

```
[3]: import pandas
      df = pandas.read_csv("bank-full.csv", sep=";")
```

```
[4]: df.head()
```

```
[4]:   age      job  marital  education  default  balance  housing  loan  \
0   58  management  married  tertiary     no     2143     yes   no
1   44  technician  single  secondary     no        29     yes   no
2   33  entrepreneur  married  secondary     no         2     yes  yes
3   47  blue-collar  married   unknown     no    1506     yes   no
4   33     unknown  single   unknown     no         1     no   no

      contact  day month  duration  campaign  pdays  previous  poutcome   y
0  unknown    5  may      261         1     -1         0  unknown  no
1  unknown    5  may      151         1     -1         0  unknown  no
2  unknown    5  may       76         1     -1         0  unknown  no
3  unknown    5  may       92         1     -1         0  unknown  no
4  unknown    5  may      198         1     -1         0  unknown  no
```

```
[5]: df.to_csv("bank_full_tab_no.txt", sep="\t", index=False, header=None)
```

1.2 Connect to the cluster

```
[6]: import pyensae
blobstorage =
blobpassword =
hadoop_server =
hadoop_password =
username = "centrale"
client, bs = %hd_open
client, bs
```

```
[6]: (<pyensae.remote.azure_connection.AzureClient at 0x1a349a00550>,
<azure.storage.blob.blockblobservice.BlockBlobService at 0x1a349a314a8>)
```

1.3 Upload the data

```
[7]: %blob_up bank_full_tab_no.txt hdblobstorage/centrale2/bank_full_tab_no.txt
```

```
[7]: 'centrale2/bank_full_tab_no.txt'
```

```
[8]: %blob_ls hdblobstorage/centrale2
```

```
[8]:
          name                last_modified content_type \
0 centrale2/bank_full_tab_no.txt 2016-06-16 10:18:58+00:00      None

      content_length blob_type
0           3751188 BlockBlob
```

1.4 Submit a PIG query

```
[9]: mapping = {'int64': 'double', 'float': 'double', 'object': 'chararray'}
schema = ["%s:%s" % (_, mapping.get(str(_[1]), _[1])) for _ in zip(df.columns, df.
↳ dtypes)]
schema = ", ".join(schema)
schema
```

```
[9]: 'age:double, job:chararray, marital:chararray, education:chararray,
default:chararray, balance:double, housing:chararray, loan:chararray,
contact:chararray, day:double, month:chararray, duration:double,
campaign:double, pdays:double, previous:double, poutcome:chararray, y:chararray'
```

On ajoute l'instruction `DESCRIBE`.

```
[10]: %%PIG_azure aggage3.pig
values = LOAD '$CONTAINER/centrale/bank_full_tab_no.txt' USING PigStorage('\t') AS
↳ (age:double,
      job:chararray, marital:chararray, education:chararray,
      default:chararray, balance:double, housing:chararray, loan:
↳ chararray,
      contact:chararray, day:double, month:chararray, duration:double,
      campaign:double, pdays:double, previous:double, poutcome:chararray,
↳ y:chararray);
DESCRIBE values;
gr = GROUP values BY loan ;
```

```
DESCRIBE gr;
agg = FOREACH gr GENERATE group, AVG(age) AS avg_age ;
DESCRIBE agg;
STORE agg INTO '$CONTAINER/centrale/bank_full_tab_no_agg.txt' USING PigStorage('\t') ;
```

```
[11]: jid = %hd_pig_submit aggage3.pig
```

```
[12]: jid
```

```
[12]: {'id': 'job_1466069083851_0005'}
```

```
[13]: %hd_queue
```

```
[13]: [{'detail': None, 'id': 'job_1466069083851_0005'},
      {'detail': None, 'id': 'job_1466069083851_0004'},
      {'detail': None, 'id': 'job_1466069083851_0003'},
      {'detail': None, 'id': 'job_1466069083851_0002'},
      {'detail': None, 'id': 'job_1466069083851_0001'}]
```

```
[14]: df = %hd_job_status jid['id']
df["status"]["state"]
```

```
[14]: 'RUNNING'
```

```
[15]: %hd_tail_stderr -n 100 jid['id']
```

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

```
[16]: %%PIG_azure aggage4.pig
values = LOAD '$CONTAINER/centrale/bank_full_tab_no.txt' USING PigStorage('\t') AS
  ↳(age:double,
                                     job:chararray, marital:chararray,
  ↳education:chararray,
                                     default:chararray, balance:double,
  ↳housing:chararray, loan:chararray,
                                     contact:chararray, day:double,
  ↳month:chararray, duration:double,
                                     campaign:double,
                                     pdays:double, previous:double,
  ↳poutcome:chararray, y:chararray);
DESCRIBE values;
gr = GROUP values BY loan ;
DESCRIBE gr;
agg = FOREACH gr GENERATE group, AVG(values.age) AS avg_age ;
DESCRIBE agg;
STORE agg INTO '$CONTAINER/centrale/bank_full_tab_no_agg2.txt' USING PigStorage('\t') ;
```

```
[17]: jid = %hd_pig_submit aggage4.pig
```

```
[18]: jid
```

```
[18]: {'id': 'job_1466069083851_0008'}
```

```
[19]: %hd_queue
```

```
[19]: [{'detail': None, 'id': 'job_1466069083851_0009'},  
      {'detail': None, 'id': 'job_1466069083851_0008'},  
      {'detail': None, 'id': 'job_1466069083851_0007'},  
      {'detail': None, 'id': 'job_1466069083851_0006'},  
      {'detail': None, 'id': 'job_1466069083851_0005'},  
      {'detail': None, 'id': 'job_1466069083851_0004'},  
      {'detail': None, 'id': 'job_1466069083851_0003'},  
      {'detail': None, 'id': 'job_1466069083851_0002'},  
      {'detail': None, 'id': 'job_1466069083851_0001'}]
```

```
[20]: df = %hd_job_status jid['id']  
df["status"]["state"]
```

```
[20]: 'RUNNING'
```

```
[21]: hd_tail_stderr -n 50 jid['id']
```

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

```
[22]: %blob_ls /centrale
```

```
[22]:
```

	name	last_modified	\
0	centrale/bank_full.csv	2016-06-15 22:17:59+00:00	
1	centrale/bank_full_tab.txt	2016-06-15 22:19:46+00:00	
2	centrale/bank_full_tab_no.txt	2016-06-15 23:00:52+00:00	
3	centrale/bank_full_tab_no_agg.txt	2016-06-16 10:32:11+00:00	
4	centrale/bank_full_tab_no_agg.txt/_SUCCESS	2016-06-16 10:32:11+00:00	
5	centrale/bank_full_tab_no_agg.txt/part-r-00000	2016-06-16 10:32:11+00:00	
6	centrale/bank_full_tab_no_agg2.txt	2016-06-16 21:13:14+00:00	
7	centrale/bank_full_tab_no_agg2.txt/_SUCCESS	2016-06-16 21:13:14+00:00	
8	centrale/bank_full_tab_no_agg2.txt/part-r-00000	2016-06-16 21:13:13+00:00	
9	centrale/scripts/pig/aggage.pig	2016-06-15 23:15:54+00:00	
10	centrale/scripts/pig/aggage.pig.log	2016-06-15 23:16:40+00:00	
11	centrale/scripts/pig/aggage.pig.log/exit	2016-06-15 23:16:40+00:00	
12	centrale/scripts/pig/aggage.pig.log/stderr	2016-06-15 23:16:30+00:00	
13	centrale/scripts/pig/aggage.pig.log/stdout	2016-06-15 23:16:30+00:00	
14	centrale/scripts/pig/aggage2.pig	2016-06-16 10:28:16+00:00	
15	centrale/scripts/pig/aggage2.pig.log	2016-06-16 10:29:04+00:00	
16	centrale/scripts/pig/aggage2.pig.log/exit	2016-06-16 10:29:04+00:00	
17	centrale/scripts/pig/aggage2.pig.log/stderr	2016-06-16 10:28:54+00:00	
18	centrale/scripts/pig/aggage2.pig.log/stdout	2016-06-16 10:28:54+00:00	
19	centrale/scripts/pig/aggage3.pig	2016-06-16 21:05:11+00:00	
20	centrale/scripts/pig/aggage3.pig.log	2016-06-16 21:05:59+00:00	
21	centrale/scripts/pig/aggage3.pig.log/exit	2016-06-16 21:05:59+00:00	
22	centrale/scripts/pig/aggage3.pig.log/stderr	2016-06-16 21:05:49+00:00	
23	centrale/scripts/pig/aggage3.pig.log/stdout	2016-06-16 21:05:49+00:00	
24	centrale/scripts/pig/aggage4.pig	2016-06-16 21:11:47+00:00	
25	centrale/scripts/pig/aggage4.pig.log	2016-06-16 21:13:31+00:00	
26	centrale/scripts/pig/aggage4.pig.log/exit	2016-06-16 21:13:31+00:00	
27	centrale/scripts/pig/aggage4.pig.log/stderr	2016-06-16 21:13:21+00:00	
28	centrale/scripts/pig/aggage4.pig.log/stdout	2016-06-16 21:13:21+00:00	
29	centrale2/bank_full_tab_no.txt	2016-06-16 10:18:58+00:00	

	content_type	content_length	blob_type
0	None	4610348	BlockBlob
1	None	3751306	BlockBlob
2	None	3751306	BlockBlob
3	None	0	BlockBlob
4	None	0	BlockBlob
5	None	49	BlockBlob
6	None	0	BlockBlob
7	None	0	BlockBlob
8	None	49	BlockBlob
9	None	782	BlockBlob
10	None	0	BlockBlob
11	None	3	BlockBlob
12	None	4060	BlockBlob
13	None	0	BlockBlob
14	None	853	BlockBlob
15	None	0	BlockBlob
16	None	3	BlockBlob
17	None	4883	BlockBlob
18	None	613	BlockBlob
19	None	853	BlockBlob
20	None	0	BlockBlob
21	None	3	BlockBlob
22	None	4883	BlockBlob
23	None	613	BlockBlob
24	None	861	BlockBlob
25	None	0	BlockBlob
26	None	3	BlockBlob
27	None	16643	BlockBlob
28	None	654	BlockBlob
29	None	3751188	BlockBlob

```
[23]: %blob_downmerge --help
```

```
usage: blob_downmerge [-h] [-o] remotepath localfile
```

download a set of files from a blob storage folder, files will be merged, we assume the container is the first element to the remote path

positional arguments:

```
remotepath      remote path of the folder to download
localfile       local name for the downloaded merged file
```

optional arguments:

```
-h, --help      show this help message and exit
-o, --overwrite overwrite the local file
```

```
usage: blob_downmerge [-h] [-o] remotepath localfile
```

```
[24]: %blob_down /centrale/bank_full_tab_no_agg2.txt/part-r-00000 agg_hadoop3.txt
```

```
[24]: 'agg_hadoop3.txt'
```

```
[25]: import pandas
df = pandas.read_csv("agg_hadoop3.txt", sep="\t", header=-1)
df
```

```
[25]:      0      1
0  no  41.008823
1  yes 40.555632
2  loan      NaN
```

J'ai oublié d'enlever le header. On vérifie que les calculs sont bons en les faisant en local.

```
[26]: df = pandas.read_csv("bank-full.csv", sep=";")
df.head()
```

```
[26]:   age      job  marital  education  default  balance  housing  loan  \
0   58  management  married  tertiary     no     2143     yes   no
1   44  technician  single  secondary     no      29     yes   no
2   33  entrepreneur  married  secondary     no      2     yes  yes
3   47  blue-collar  married  unknown     no    1506     yes   no
4   33      unknown  single  unknown     no      1     no   no

   contact  day month  duration  campaign  pdays  previous  poutcome  y
0  unknown    5  may     261         1     -1         0  unknown  no
1  unknown    5  may     151         1     -1         0  unknown  no
2  unknown    5  may      76         1     -1         0  unknown  no
3  unknown    5  may      92         1     -1         0  unknown  no
4  unknown    5  may     198         1     -1         0  unknown  no
```

```
[27]: df[["loan", "age"]].groupby("loan").mean()
```

```
[27]:      age
loan
no    41.008823
yes   40.555632
```

```
[28]:
```