pig_params_azure_correction

July 1, 2023

1 PIG et Paramètres (Azure) (correction)

Correction.

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

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

1.1 Connexion au cluster

On prend le cluster Cloudera. Il faut exécuter ce script pour pouvoir notifier au notebook que la variable `params` existe.

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

[2]:
```
<IPython.core.display.HTML at 0x6c9f270>
```

[3]:
```
import pyensae
%load_ext pyensae
%load_ext pyenbc
blobstorage = blobhp["blob_storage"]
blobpassword = blobhp["password1"]
hadoop_server = blobhp["hadoop_server"]
hadoop_password = blobhp["password2"]
username = blobhp["username"]
client, bs =  %hd_open
client, bs
```

[3]:
```
(<pyensae.remote.azure_connection.AzureClient at 0x99f4b10>,
 <azure.storage.blobservice.BlobService at 0x99f4b50>)
```

1.2 Exercice 1 : min, max

On ajoute deux paramètres pour construire l'histogramme entre deux valeurs `a`,`b`. Ajouter ces deux paramètres au nom du fichier de sortie peut paraître raisonnable mais l'interpréteur a du mal à identifier les paramètres `Undefined parameter : bins_`. On utilise des tirets.

[4]:
```
%%PIG histogramab.pig

values = LOAD '$CONTAINER/$PSEUDO/random/random.sample.txt' USING PigStorage('\t') AS␣
↪(x:double);

values_f = FILTER values BY x >= $a AND x <= $b ;   -- ligne ajoutée

values_h = FOREACH values_f GENERATE x, ((int)(x / $bins)) * $bins AS h ;

hist_group = GROUP values_h BY h ;

hist = FOREACH hist_group GENERATE group, COUNT(values_h) AS nb ;

STORE hist INTO '$CONTAINER/$PSEUDO/random/histo_$bins-$a-$b.txt' USING␣
↪PigStorage('\t') ;
```

[5]:
```
if client.exists(bs, client.account_name, "$PSEUDO/random/histo_0.1-0.2-0.8.txt"):
    r = client.delete_folder (bs, client.account_name, "$PSEUDO/random/histo_0.1-0.2-0.
↪8.txt")
    print(r)
```

[6]:
```
jid = client.pig_submit(bs, client.account_name, "histogramab.pig",
                   params = dict(bins="0.1", a="0.2", b="0.8") )
jid
```

[6]: {'id': 'job_1416874839254_0202'}

[7]:
```
st = %hd_job_status jid["id"]
st["id"],st["percentComplete"],st["status"]["jobComplete"]
```

[7]: ('job_1416874839254_0202', '100% complete', True)

[8]:
```
%hd_tail_stderr jid["id"]
```

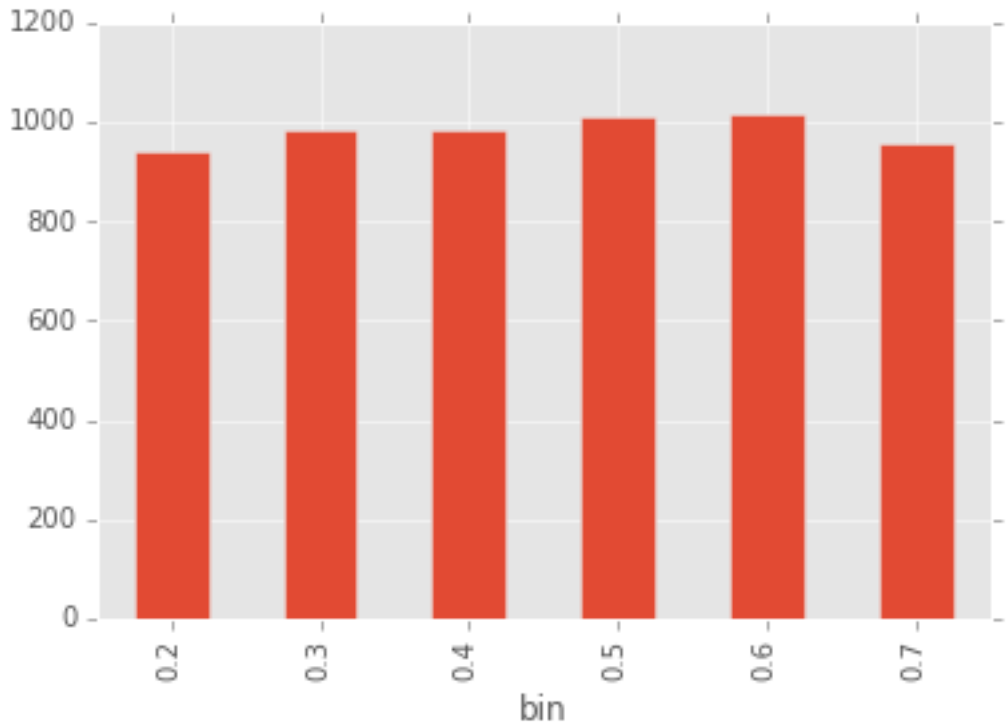
[8]: <IPython.core.display.HTML at 0x9bc24f0>

[9]:
```
import os
if os.path.exists("histo.txt") : os.remove("histo.txt")
%blob_downmerge /$PSEUDO/random/histo_0.1-0.2-0.8.txt histo.txt
```

[9]: 'histo.txt'

[10]:
```
import matplotlib.pyplot as plt
plt.style.use('ggplot')
import pandas
df = pandas.read_csv("histo.txt", sep="\t",names=["bin","nb"])
df.plot(x="bin",y="nb",kind="bar")
```

[10]: <matplotlib.axes._subplots.AxesSubplot at 0xb0874d0>



[11]: