如何从pyspark中的hdfs获取目录的文件名列表?

bnemo (https://oomake.com/people/bnemo) 发布于 2019-11-01 • 在 directory (/topic/directory) • 最后更新 2019-11-01 19:32 • 659 浏览

我在hdfs中有一个包含许多文件的目录。我知道目录的路径，我试图获取目录包含的文件名列表。我怎么能这样做？ 如果我有一个目录如下：

```
+dir/
    +f1
    +f2
    +fN
```

我想获得如下列表：

```
[f1, f2, fN]
```

Pyspark中没有这方面的功能(编辑：最后请参阅Mariusz和UPDATE的回答) - 此功能在Python包pywebhdfs (https://pypi.python.org/pypi/pywebhdfs)中提供(只需通过 `pip install pywebhdfs` 安装)：

```
from pywebhdfs.webhdfs import PyWebHdfsClient
from pprint import pprint
hdfs = PyWebHdfsClient(host='192.10.10.73',port='50070', user_name='ctsats')  # your Nameno
de IP & username here
my_dir = 'user/ctsats'
pprint(hdfs.list_dir(my_dir))
```

结果是一个(相当长的)Python字典(未显示) - 尝试一点以获得感觉。您可以解析它以获取名称和类型(文件/目录)，如下所示：

```
data = hdfs.list_dir(my_dir)
dd = [[x["pathSuffix"], x["type"]] for x in data["FileStatuses"]["FileStatus"]]
dd
# [[u'.Trash', u'DIRECTORY'], [u'.sparkStaging', u'DIRECTORY'], [u'checkpoint', u'DIRECTOR
Y'], [u'datathon', u'DIRECTORY'], [u'ms-spark', u'DIRECTORY'], [u'projects', u'DIRECTORY'],
[u'recsys', u'DIRECTORY'], [u'sparklyr', u'DIRECTORY'], [u'test.data', u'FILE'], [u'word2ve
c', u'DIRECTORY']]
```

从这里，一个简单的列表理解应该做的工作 - 例如，在我的情况下，我有两个文件&目录存在，这里是我如何只保留目录：

```
sub_dirs = [x[0] for x in dd if x[1]=='DIRECTORY']
sub_dirs
# [u'.Trash', u'.sparkStaging', u'checkpoint', u'datathon', u'ms-spark', u'projects', u'rec
sys', u'sparklyr', u'word2vec']
```

为了比较，这里是同一目录的实际列表：

```
[ctsats@dev-hd-01 ~]$ hadoop fs -ls
Found 10 items
drwx------   - ctsats supergroup          0 2016-06-08 13:31 .Trash
drwxr-xr-x   - ctsats supergroup          0 2016-12-15 20:18 .sparkStaging
drwxr-xr-x   - ctsats supergroup          0 2016-06-23 13:23 checkpoint
drwxr-xr-x   - ctsats supergroup          0 2016-02-03 15:40 datathon
drwxr-xr-x   - ctsats supergroup          0 2016-04-25 10:56 ms-spark
drwxr-xr-x   - ctsats supergroup          0 2016-06-30 15:51 projects
drwxr-xr-x   - ctsats supergroup          0 2016-04-14 18:55 recsys
drwxr-xr-x   - ctsats supergroup          0 2016-11-07 12:46 sparklyr
-rw-r--r--   3 ctsats supergroup         90 2016-02-03 16:55 test.data
drwxr-xr-x   - ctsats supergroup          0 2016-12-15 20:18 word2vec
```

必须启用Hadoop集群中的WebHDFS服务，即您的 `hdfs-site.xml` 文件必须包含以下条目：

```
<property>
    <name>dfs.webhdfs.enabled</name>
    <value>true</value>
</property>
```

更新(在Mariusz的回答之后)：这是Mariusz对Spark 1.6的回答(你需要用 `sc` 替换 `spark`)的改编：

```
path="/user/ctsats"
fs = sc._jvm.org.apache.hadoop.fs.FileSystem.get(sc._jsc.hadoopConfiguration())
list_status = fs.listStatus(sc._jvm.org.apache.hadoop.fs.Path(path))
result = [file.getPath().getName() for file in list_status]
result
# [u'.Trash', u'.sparkStaging', u'checkpoint', u'datathon', u'ms-spark', u'projects', u'rec
sys', u'sparklyr', u'test.data', u'word2vec']
```

这里的问题是返回文件和子文件夹，没有任何区分它们的方法。正如所示， `pywebhdfs` 解决方案不会受此影响..... 我想有办法克服这个问题，但你必须深入研究py4j API - 尽管有欺骗性的外观， `list_status` 不是Python列表：

```
list_status
# JavaObject id=o40
```


您可以在pyspark中使用HDFS(或任何其他兼容的Hadoop文件系统)API，并使用一些py4j魔法。要列出特定目录中的文件，请使用：

```
path = "/here/is/my/dir/"
fs = spark._jvm.org.apache.hadoop.fs.FileSystem.get(spark._jsc.hadoopConfiguration())
list_status = fs.listStatus(spark._jvm.org.apache.hadoop.fs.Path(path))
result = [file.getPath().getName() for file in list_status]
```

`list_status` 集合的元素属于FileSystem (https://hadoop.apache.org/docs/current/api/org/apache/hadoop/fs/FileStatus.html)类型。使用此API，您可以获取文件元数据，例如信息，如果它是目录，模式，所有者，组，acls，并使用这些信息来过滤掉不需要的文件。

0 0 分享 2019-11-01

list_status不需要是Python列表。它显然是一个迭代器，这就是你所需要的。如果你真的想要一个python列表，那么很容易制作一个：

```
pythonlist = list(list_status)
Other than that, you can filter out the files, and omit the directories:
result = [file.getPath().getName() for file in list_status if file.isFile()]
```

翻转

0 0 分享 2019-11-01