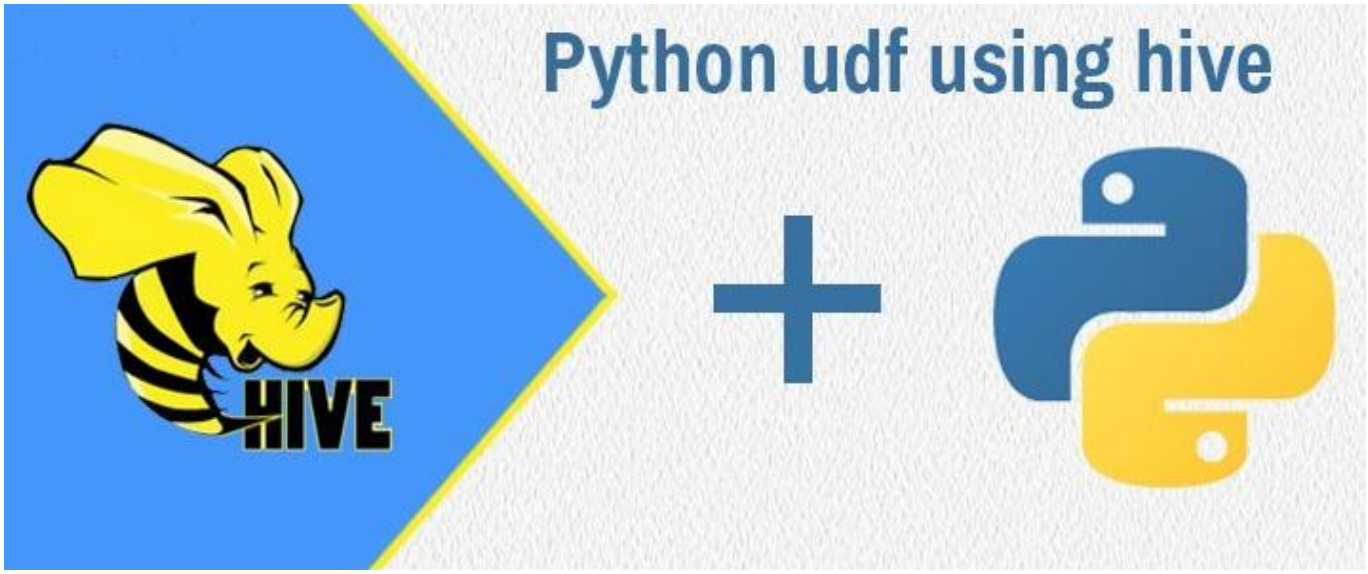
使用 Python 编写 Hive UDF 环境问题

在 《使用Python编写Hive UDF》 文章中，我简单的谈到了如何使用 Python 编写 Hive UDF 解决实际的问题。我们那个例子里面仅仅是一个很简单的示例，里面仅仅引入了 Python 的 sys 包，而这个包是 Python 内置的，所有我们不需要担心 Hadoop 集群中的 Python 没有这个包；但是问题来了，如果我们现在需要使用到 numpy 中的一些函数呢？假设我们有一个表，表结构和数据如下：

```
hive (iteblog)> show create table test_a;
OK
CREATE TABLE `test_a`(
  `data` string
)

hive (iteblog)> select * from test_a;
OK
1,3,4,5
2,5,6,6
3,5,6,7
7,1,5,8
Time taken: 0.321 seconds, Fetched: 4 row(s)
```



如果想及时了解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

每行数据是一个数字序列，现在我们需要求每行数字的最小值，我们使用 Python 编写的代码如下：

```
#!/usr/bin/python

import sys
import numpy as np

def toInt(num):
  return int(num)

for line in sys.stdin:
        x = line.split(",")
        v = np.array(map(toInt, x))
        print v.min()
```

这里面使用了 numpy 相关的函数，现在我们来使用这个 UDF：

```
hive (iteblog)> add file /tmp/iteblog.py;
Added resources: [/tmp/iteblog.py]
hive (iteblog)>select
            >    TRANSFORM(data)
            >    USING 'python iteblog.py'
            >     as (min_num)
            >from test_a;
```

不幸的是，我们的程序运行遇到了问题：

```
Traceback (most recent call last):
  File "iteblog.py", line 4, in <module>
    import numpy as np
ImportError: No module named numpy

org.apache.hadoop.hive.ql.metadata.HiveException: [Error 20003]: An error occurred when try
ing to close the Operator running your custom script.
  at org.apache.hadoop.hive.ql.exec.ScriptOperator.close(ScriptOperator.java:560)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.mr.ExecMapper.close(ExecMapper.java:192)
```

```
  at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
  at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:429)
  at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
  at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
  at java.security.AccessController.doPrivileged(Native Method)
  at javax.security.auth.Subject.doAs(Subject.java:422)
  at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491
)
  at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)
org.apache.hadoop.hive.ql.metadata.HiveException: [Error 20003]: An error occurred when try
ing to close the Operator running your custom script.
  at org.apache.hadoop.hive.ql.exec.ScriptOperator.close(ScriptOperator.java:560)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.mr.ExecMapper.close(ExecMapper.java:192)
  at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
  at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:429)
  at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
  at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
  at java.security.AccessController.doPrivileged(Native Method)
  at javax.security.auth.Subject.doAs(Subject.java:422)
  at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491
)
  at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)
org.apache.hadoop.hive.ql.metadata.HiveException: [Error 20003]: An error occurred when try
ing to close the Operator running your custom script.
  at org.apache.hadoop.hive.ql.exec.ScriptOperator.close(ScriptOperator.java:560)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.Operator.close(Operator.java:630)
  at org.apache.hadoop.hive.ql.exec.mr.ExecMapper.close(ExecMapper.java:192)
  at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:61)
  at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:429)
  at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
  at org.apache.hadoop.mapred.YarnChild$2.run(YarnChild.java:162)
  at java.security.AccessController.doPrivileged(Native Method)
  at javax.security.auth.Subject.doAs(Subject.java:422)
  at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491
)
  at org.apache.hadoop.mapred.YarnChild.main(YarnChild.java:157)
```

从错误中我们可以看出，集群的 Python 环境并没有 numpy 的环境，所有出现了 No module

named numpy 的异常。这时候我们可以通知集群的维护人员给我们部署好相关的环境，但是这个可能很麻烦。不过高兴的是，我们其实可以自己部署相关的环境，操作如下：

```
hive (iteblog)> add ARCHIVE /home/iteblog/anaconda2.tar.gz;
hive (iteblog)> add file /tmp/iteblog.py;
hive (iteblog)> select
              >   TRANSFORM(data)
              >   USING 'anaconda2.tar.gz/anaconda2/bin/python iteblog.py'
              >   as (min_num)
              > from test_a;
OK
1
2
3
1
Time taken: 32.728 seconds, Fetched: 4 row(s)
```

这次我们顺利的解决了这个问题。注意，本文提供的方法只是一种可行的方案，并不是推荐大家都这么使用，正如本文例子里面用到的 anaconda2 包，它的大小有 1G 多，如果大家都这么使用势必造成网络带宽的问题，所有如果这个包真的很常用，还是得和集群管理员商量部署好。