

**My project:** writing UDF for Hive in python

**Aim :** mask personally identifiable information (PII) data on Cloudera Hadoop Quickstart vm.

**Skills :** sql, hive, hive UDF, python, command line, etc

### 1. Create a data set in lower case.

```
[cloudera@quickstart Desktop]$ cat mycontact.txt
john    smith    415-222-3333    johns@gmail.com
sob     smith    408-222-3334    bobs@gmail.com
mark    taylor    510-222-3335    markt@gmail.com
pat     taylor    650-222-3336    patt@gmail.com
[cloudera@quickstart Desktop]$
```

### 2. Create a Table in Hive

```
CREATE table mycontact(fname STRING, lname STRING, phone STRING, email STRING)
```

row format delimited fields terminated by '\t' stored as textfile;

```
hive> create table mycontact(fname string, lname string, phone string, email string)
> row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 0.71 seconds
hive>
```

### 3. Load the Data

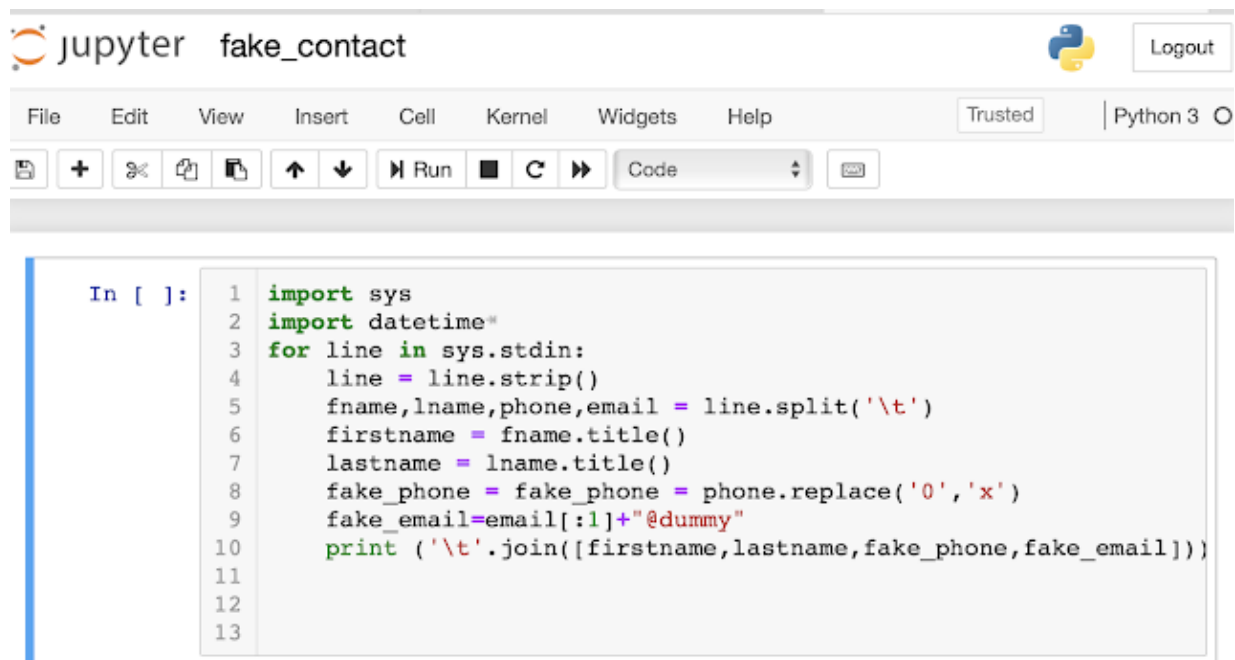
load data local inpath '/home/cloudera/Desktop/mycontact.txt' overwrite into table mycontact;

```
hive> load data local inpath '/home/cloudera/Desktop/mycontact.txt' overwrite in
to table mycontact;
Loading data to table default.mycontact
Table default.mycontact stats: [numFiles=1, numRows=0, totalSize=160, rawDataSiz
e=0]
OK
Time taken: 0.918 seconds
hive> select * from mycontact;
OK
john    smith    415-222-3333    johns@gmail.com
sob     smith    408-222-3334    bobs@gmail.com
mark    taylor    510-222-3335    markt@gmail.com
pat     taylor    650-222-3336    patt@gmail.com
Time taken: 0.499 seconds, Fetched: 4 row(s)
```

### 4. Write UDF in Python

```
import sys
import datetime
for line sys.stdin: # Read data from STDIN.
    line = line.strip() # Strip the data into individual lines.
    fname,lname,phone,email = line.split('\t') # Split the line into words based on
tab delimiter
    firstname = fname.title() # covert all first characters into upper case
    lastname = lname.title()
    fake_phone = fake_phone = phone.replace('0','x') #convert phone number '0'
into 'x'
    fake_email=fake_email=email[:1]+"@dummy.com" #convert email address ending
@dummy
```

```
print ('\t'.join([firstname,lastname,fake_phone,fake_email]))
```



The image shows a Jupyter Notebook interface with the title 'fake\_contact'. The top bar includes the Jupyter logo, the title, a Python logo, and a 'Logout' button. Below the top bar is a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. To the right of the menu bar are 'Trusted' and 'Python 3' indicators. Below the menu bar is a toolbar with icons for saving, adding cells, zooming, copying, pasting, undo, redo, and running code. The main area contains a code cell with the following Python code:

```
In [ ]: 1 import sys
        2 import datetime
        3 for line in sys.stdin:
        4     line = line.strip()
        5     fname,lname,phone,email = line.split('\t')
        6     firstname = fname.title()
        7     lastname = lname.title()
        8     fake_phone = fake_phone = phone.replace('0','x')
        9     fake_email=email[:1]+"@dummy"
       10     print ('\t'.join([firstname,lastname,fake_phone,fake_email]))
       11
       12
       13
```

## 5. Add the Python File into Hive

add file /home/cloudera/Desktop/fake\_contact.py;

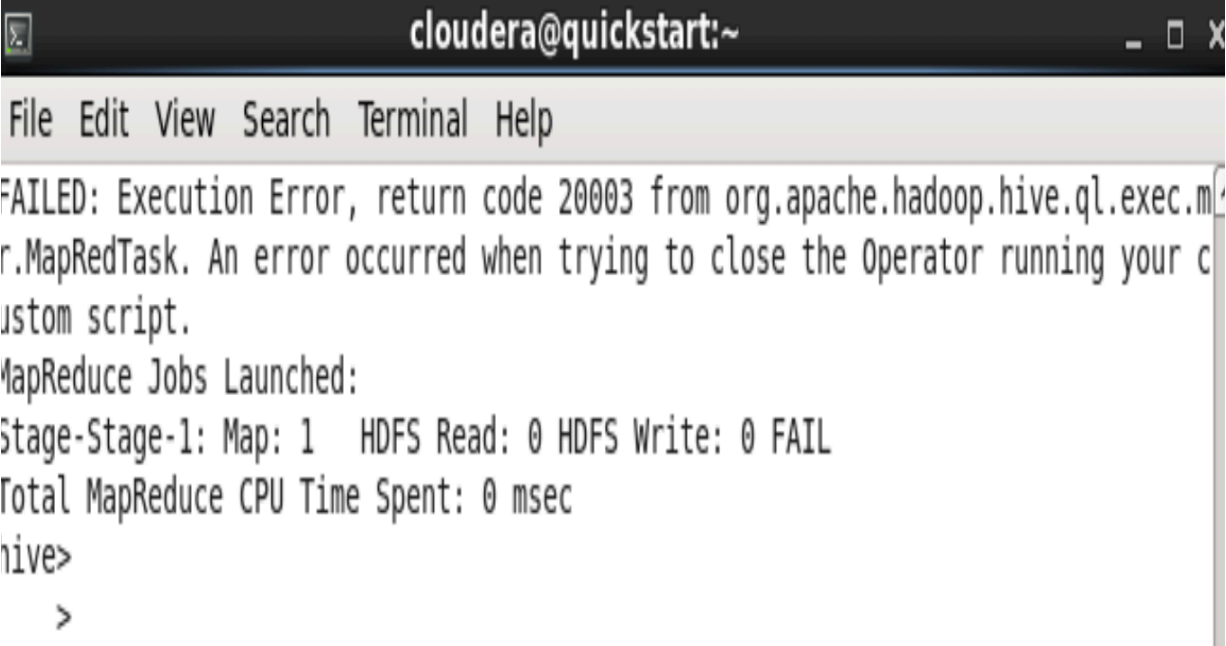
```
hive> add file /home/cloudera/Desktop/fake_contact.py;
Added resources: [/home/cloudera/Desktop/fake_contact.py]
hive> █
```

## 6. Run the Python UDF using 'transform'

```
SELECT TRANSFORM (fname,lname,phone,email) USING 'python fake_contact.py'
AS (firstname,lastname,fake_phone,fake_email) FROM mycontact;
```

```
hive> select transform (fname,lname,phone,email)
> using 'python fake_contact.py'
> as (firstname,lastname,fake_phone,fake_email)
> from mycontact;
Query ID = cloudera_20190823132929_050f67e4-dd3d-41d4-bc51-16aa03618266
Total jobs = 1
```

7. Fail: Execution error since I forget one "(" )" in python code.

A terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal output shows a Hive query failure. The error message is: "FAILED: Execution Error, return code 20003 from org.apache.hadoop.hive.ql.exec.m...r.MapRedTask. An error occurred when trying to close the Operator running your c...ustom script." Below this, it says "MapReduce Jobs Launched:" followed by "Stage-Stage-1: Map: 1 HDFS Read: 0 HDFS Write: 0 FAIL" and "Total MapReduce CPU Time Spent: 0 msec". The prompt "hive>" is shown again with a cursor on the next line.

```
cloudera@quickstart:~
File Edit View Search Terminal Help
FAILED: Execution Error, return code 20003 from org.apache.hadoop.hive.ql.exec.m
r.MapRedTask. An error occurred when trying to close the Operator running your c
ustom script.
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 HDFS Read: 0 HDFS Write: 0 FAIL
Total MapReduce CPU Time Spent: 0 msec
hive>
>
```

8. Fix and rerun

```
MapReduce Total cumulative CPU time: 950 msec
Ended Job = job_1566575367058_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 0.95 sec HDFS Read: 4753 HDFS Write: 144 SUCCESS
Total MapReduce CPU Time Spent: 950 msec
OK
John    Smith    415-222-3333    j@dummy.com
Sob     Smith    4x8-222-3334    b@dummy.com
Mark    Taylor    51x-222-3335    m@dummy.com
Pat     Taylor    65x-222-3336    p@dummy.com
Time taken: 15.403 seconds, Fetched: 4 row(s)
hive> █
```

## 8. Result

name : john smith -> John Smith  
phone number : 510-222-3355 -> 5x0-222-3355  
email: johns@gmail.com -> j@dummy.com