**Big Data / ETL Engineer Coding exercise**

Instructions: Submit the code either in your git hub public repository and send the URL or share the zipped “.py” files

1. Create a python functionality to do logging to hdfs

given info

hdfs\_name\_node\_ip='x.x.x.x'

hdfs\_name\_ndoe\_port='1234'

hdfs\_user\_name='test'

file\_path=’/user/logger/python/logs’

file\_name=’test.log’

1. Create a hive context in pyspark and set the below properties

hive.merge.mapredfiles=false

hive.merge.smallfiles.avgsize=16000000

hive.execution.engine=mr

1. Assume there is a column called email in a dataframe

create a code snippet to generate a new column email\_valid\_flag which flags if the email is valid or not

in pandas

in pyspark

(write your own email validator function don’t assume or import)

1. Convert a dataframe to list of dicts and read the same into a another dataframe with prefix of "a\_" apended to the original column name (assume some data)
2. Code using pandas
3. Code using pyspark
4. Using pyspark, create snippet to generate a sequential id(*should be consecutive numbers*) column in a dataframe
5. Create a  sample smartsheet using below link

(<https://www.smartsheet.com/(need> to create a account first)

Read the data from this smartsheet using its api and post it to  AWS bucket

with bucketname='test\_bucket'(note the bucket has server-side AES256 encryption enabled)

key='test\_key'

using python (*do it without creating local files and preferably doing a raw request and parsing the returned json***)**

7) Write a code snippet to do pivot table of a dataframe using pandas(assume some data) without using pandas pivot\_table functions

8) Write a code snippet to read a table from sqlitedb  into pandas dataframe(assume some data)

9) Assume that there is directory with mix of csv files delimited by comma(,),tab(\t) and pipe(|)

they all have the same number of columns.

create a python code snippet to read the files from above directory

*print the filename and delimiter used in the file*

into a *single* pandas dataframe

and create a spark dataframe of the above pandas dataframe

10)

There is pyspark dataframe with a column named city.

Create a snippet to add a new column called country

Don’t use any python libraries but you are allowed to read csv file with country and city mappings.