This is the code to merge and compress JSONs on AWS S3.

Number of JSONs = Number of spark executors

The ‘array’ contains names of folders to iterate on S3

How to compress JSONS ->

1. Load all JSONs into dataframe (df)
2. df. write.format("json").save(saveLocation)
3. To combine compress even more (n JSONs)
4. df.coalesce(n).write.format("json").save(saveLocation)

----------------------------------------Code to compress all folders-------------------------------------

%pyspark

root = "s3n://data-lake-us-east-2-549323063936-encrypted/Agrian/AdHoc/NoETL/"

array = ['classification', 'commodity', 'commodity\_variety', 'country', "county", 'crop\_plan', 'crop\_plan\_application\_activity', 'crop\_plan\_blueprint', 'crop\_plan\_category', 'crop\_plan\_combined\_activity', 'crop\_plan\_custom\_phase\_set', 'crop\_plan\_generic\_activity', 'crop\_plan\_phase', 'crop\_plan\_planting\_activity', 'crop\_plan\_scenario', 'crop\_plan\_simple\_activity', 'crop\_plan\_sub\_phase', 'event', 'farm', 'farm\_water\_source', 'field', 'field\_event', 'field\_event\_layer', 'grower', 'grower\_method', 'ingredient', 'intended\_use', 'irrigation\_source', 'list', 'material', 'note', 'organization', 'pest', 'planting', 'planting\_method', 'product', 'sample\_event', 'soil\_order', 'soil\_texture', 'state', 'tillage\_type', 'trap']

arrayMin = [ 'farm', 'field', 'grower', 'sample\_event' ]

for data\_field in arrayMin:

rootLocation = root + data\_field + "/\*.json"

print saveLocation + data\_field + "\_exploded"

try:

print "Saving " + data\_field

df = sqlContext.read.json(rootLocation)

#newDf = relationalize(df)

# relationalize function is from flattenJSONPySpark

newDf.coalesce(1).write.format("json").save(saveLocation + "exploded" + data\_field + ".json")

except:

print "Something went wrong with saving " + data\_field

----------------------------------------Code to compress all folders-------------------------------------