MVP TEST INSTRUCTION

SETUP ENVIRONMENT(LOCAL):

1. Download Spark from :http://spark.apache.org/downloads.html (version:1.5.2)
   please choose the approriate package type according to Hadoop version.

2. To build Spark and its example programs, run: build/mvn

3. Install python 2.7(or Python 3)

RUN ON CLUSTER:

1. Login Cluster:

```
    ssh honeycomb@128.2.7.38 (password: ask teammates)
```

2. Copy files into CLuster local host:

```
   sudo  scp source_file_name honeycomb@128.2.7.38:/home/honeycomb/SparkTeam
   e.g:
  sudo  scp /Users/jacobliu/PySpark.py honeycomb@128.2.7.38:/home/honeycomb/SparkTeam
```

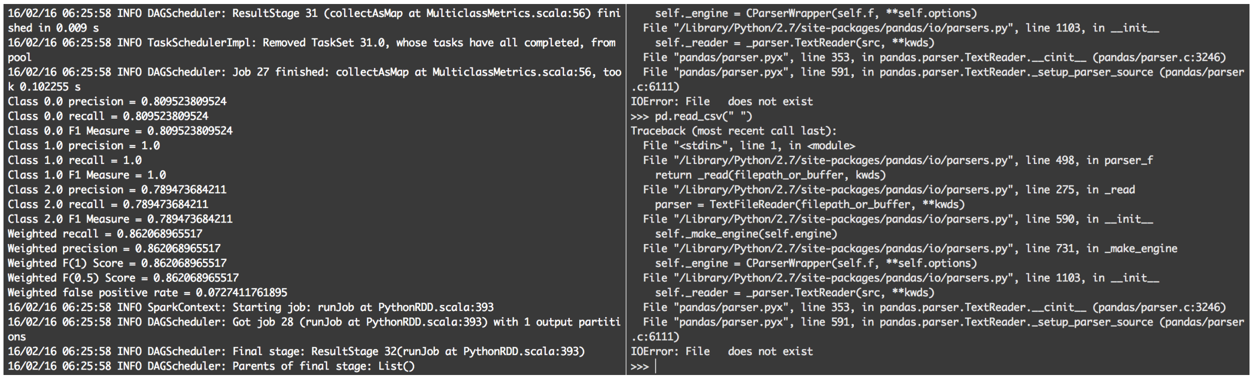
3. Put files into HDFS:

```
    HADOOP_USER_NAME=hdfs hdfs dfs -put LOCAL_FILE_PATH HDFS_FILE_PATH
    e.g:
    hdfs dfs -put /home/honeycomb/SparkTeam/sample_multiclass_classification_data_test.txt
/user/spark/input
```

4. Put PySpark.py and train/test dataset into HDFS and run command line:

```
   YOUR_SEVER_SPARK_PATH/spark-submit PySpark.py YOUR_TRAIN_DATA_HDFS_PATH
YOUT_TEST_DATA_HDFS_PATH YOUR_OUTPUT_HDFS_PATH
   e.g:
  /bin/spark-submit /home/honeycomb/SparkTeam/PySpark.py
/user/spark/input/sample_multiclass_classification_data.txt
/user/spark/input/sample_multiclass_classification_data_test.txt  /user/spark/out/
```

RESULT: SUCCESS!



```
16/02/16 06:25:58 INFO DAGScheduler: ResultStage 31 (collectAsMap at MulticlassMetrics.scala:56) fini
shed in 0.009 s
16/02/16 06:25:58 INFO TaskSchedulerImpl: Removed TaskSet 31.0, whose tasks have all completed, from
pool
16/02/16 06:25:58 INFO DAGScheduler: Job 27 finished: collectAsMap at MulticlassMetrics.scala:56, too
k 0.102255 s
Class 0.0 precision = 0.809523809524
Class 0.0 recall = 0.809523809524
Class 0.0 F1 Measure = 0.809523809524
Class 1.0 precision = 1.0
Class 1.0 recall = 1.0
Class 1.0 F1 Measure = 1.0
Class 2.0 precision = 0.789473684211
Class 2.0 recall = 0.789473684211
Class 2.0 F1 Measure = 0.789473684211
Weighted recall = 0.862068965517
Weighted precision = 0.862068965517
Weighted F(1) Score = 0.862068965517
Weighted F(0.5) Score = 0.862068965517
Weighted false positive rate = 0.0727411761895
16/02/16 06:25:58 INFO SparkContext: Starting job: runJob at PythonRDD.scala:393
16/02/16 06:25:58 INFO DAGScheduler: Got job 28 (runJob at PythonRDD.scala:393) with 1 output partiti
ons
16/02/16 06:25:58 INFO DAGScheduler: Final stage: ResultStage 32(runJob at PythonRDD.scala:393)
16/02/16 06:25:58 INFO DAGScheduler: Parents of final stage: List()
```

```
    self._engine = CParserWrapper(self.f, **self.options)
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 1103, in __init__
    self._reader = _parser.TextReader(src, **kwds)
  File "pandas/parser.pyx", line 353, in pandas.parser.TextReader.__cinit__ (pandas/parser.c:3246)
  File "pandas/parser.pyx", line 591, in pandas.parser.TextReader._setup_parser_source (pandas/parser
.c:6111)
IOError: File   does not exist
>>> pd.read_csv(" ")
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 498, in parser_f
    return _read(filepath_or_buffer, kwds)
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 275, in _read
    parser = TextFileReader(filepath_or_buffer, **kwds)
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 590, in __init__
    self._make_engine(self.engine)
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 731, in _make_engine
    self._engine = CParserWrapper(self.f, **self.options)
  File "/Library/Python/2.7/site-packages/pandas/io/parsers.py", line 1103, in __init__
    self._reader = _parser.TextReader(src, **kwds)
  File "pandas/parser.pyx", line 353, in pandas.parser.TextReader.__cinit__ (pandas/parser.c:3246)
  File "pandas/parser.pyx", line 591, in pandas.parser.TextReader._setup_parser_source (pandas/parser
.c:6111)
IOError: File   does not exist
>>> 
```