Spark Code Skeleton

Spark + RDDs

from pyspark import SparkConf, SparkContext

import sys

assert sys.version\_info >= (3, 5) # make sure we have Python 3.5+

# add more functions as necessary

def main(inputs, output):

# main logic starts here

if \_\_name\_\_ == '\_\_main\_\_':

conf = SparkConf().setAppName('example code')

sc = SparkContext(conf=conf)

# sc.setLogLevel('WARN')

assert sc.version >= '2.3' # make sure we have Spark 2.3+

inputs = sys.argv[1]

output = sys.argv[2]

main(inputs, output)

Spark + DataFrames

import sys

assert sys.version\_info >= (3, 5) # make sure we have Python 3.5+

from pyspark.sql import SparkSession, functions, types

spark = SparkSession.builder.appName('example code').getOrCreate()

assert spark.version >= '2.3' # make sure we have Spark 2.3+

# spark.sparkContext.setLogLevel('WARN')

# sc = spark.sparkContext

# add more functions as necessary

def main(inputs, output):

# main logic starts here

if \_\_name\_\_ == '\_\_main\_\_':

inputs = sys.argv[1]

output = sys.argv[2]

main(inputs, output)