Spark

《Spark快速大数据分析》

install spark

$wget <http://d3kbcqa49mib13.cloudfront.net/spark-2.0.0-bin-hadoop2.7.tgz>

$tar -xf spark…tgz

$cd spark…

交互式分析数据

启动Spark中Python的shell

$bin/pyspark

// actions, return values

lines = sc.textFile("README.md")

lines.count()

lines.first()

//transformation, return a new RDD

pythonlines = lines.filter(lamba line: "Python" in line)

textFile.map(lambda line: len(line)).reduce(lambda a,b: a if (a>b) else b)

wordCounts = textFile.flatMap(lambda line: line.split()).

map(lambda word: (word, 1)).

reduceByKey(lambda a, b: a+b)

wordCounts.collect()

独立应用

pi.py

from random import random

from pyspark import SparkContext

def sample(p):

x, y = random(), random()

return 1 if x\*x + y\*y < 1 else 0

NUM\_SAMPLES = 100000

sc = SparkContext("local", "Pi App")

count = sc.parallelize(xrange(0, NUM\_SAMPLES)).map(sample).reduce(lambda a,b: a+b)

print "Pi is roughly %f" % (4.0 \* count / NUM\_SAMPLES)

$bin/spark-submit pi.py

查看spark自带的范例

$tree examples/src/main/python

$cat examples/src/main/python/pi.py

$bin/spark-submit examples/src/main/python/pi.py

RDD转换操作

rdd.map(func)

rdd.flapMap(func)

rdd.filter(func)

rdd.distinct(func) 去重

rdd.sample(withReplacement, fraction, [seed]) 采样

rdd.union(other)

rdd.intersection(other)

rdd.subtract(other)

rdd.cartesian(other) 笛卡尔积

RDD行动操作

rdd.collect()

rdd.count()

rdd.countByValue()

rdd.take(num)

rdd.top(num)

rdd.takeOrdered(num)(myOrdering)

rdd.takeSample(withReplacement, num, [seed])

rdd.reduce(func)

rdd.fold(initValue)(func)

rdd.aggregate(initValue)(seqOp, combOp)

rdd.foreach(func)