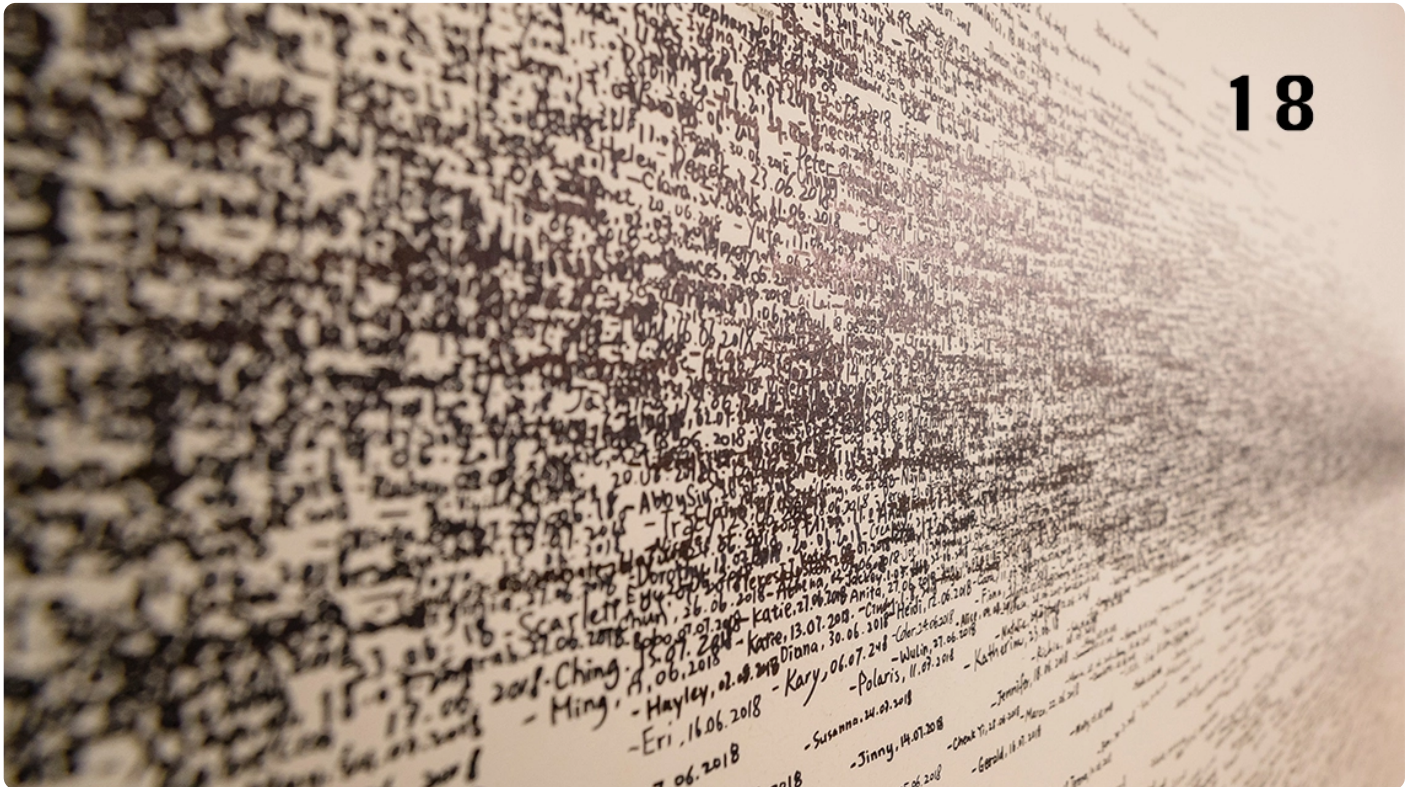


18 | Word Count: 从零开始运行你的第一个Spark应用

2019-05-29 蔡元楠 来自北京

《大规模数据处理实战》



你好，我是蔡元楠。

今天我们来从零开始运行你的第一个 Spark 应用。

我们先来回顾一下模块三的学习路径。

首先，我们由浅入深地学习了 Spark 的基本数据结构 RDD，了解了它这样设计的原因，以及它所支持的 API。

之后，我们又学习了 Spark SQL 的 DataSet/DataFrame API，了解到它不仅提供类似于 SQL query 的接口，大大提高了开发者的工作效率，还集成了 Catalyst 优化器，可以提升程序的性能。

这些 API 应对的都是批处理的场景。

再之后，我们学习了 Spark 的流处理模块：Spark Streaming 和 Structured Streaming。两者都是基于微批处理（Micro batch processing）的思想，将流数据按时间间隔分割成小的数据块进行批处理，实时更新计算结果。

其中 Structured Streaming 也是使用 DataSet/DataFrame API，这套 API 在某种程度上统一了批处理和流处理，是当前 Spark 最流行的工具，我们必需要好好掌握。

虽然学习了这么多 API 以及它们的应用，但是大部分同学还没有从零开始写一个完整的 Spark 程序，可能更没有运行 Spark 程序的经历。纸上谈兵并不能帮助我们在工作中用 Spark 解决实际问题。所以，今天我就和你一起做个小练习，从在本地安装 Spark、配置环境开始，为你示范怎样一步步解决之前提到数次的统计词频（Word Count）的问题。

通过今天的学习，你可以收获：

- 怎样安装 Spark 以及其他相关的模块；

- 知道什么是 SparkContext、SparkSession；

- 一个完整的 Spark 程序应该包含哪些东西；

- 用 RDD、DataFrame、Spark Streaming 如何实现统计词频。

这一讲中，我们使用的编程语言是 Python，操作系统是 Mac OS X。

在这一讲以及之前文章的例子中，我们都是用 Python 作为开发语言。虽然原生的 Spark 是用 Scala 实现，但是在大数据处理领域中，我个人最喜欢的语言是 Python。因为它非常简单易用，应用非常广泛，有很多的库可以方便我们开发。

当然 Scala 也很棒，作为一个函数式编程语言，它很容易用链式表达对数据集进行各种处理，而且它的运行速度是最快的，感兴趣的同学可以去学习一下。

虽然 Spark 还支持 Java 和 R，但是我个人不推荐你使用。用 Java 写程序实在有些冗长，而且速度上没有优势。


操作系统选 Mac OS X 是因为我个人喜欢使用 Macbook，当然 Linux/Ubuntu 也很棒。

安装 Spark

首先，我们来简单介绍一下如何在本地安装 Spark，以及用 Python 实现的 Spark 库——PySpark。

在前面的文章中，我们了解过，Spark 的 job 都是 JVM (Java Virtual Machine) 的进程，所以在安装运行 Spark 之前，我们需要确保已经安装 Java Developer Kit (JDK)。在命令行终端中输入：

```
1 java -version
```

 复制代码

如果命令行输出了某个 Java 的版本，那么说明你已经有 JDK 或者 JRE 在本地。如果显示无法识别这个命令，那么说明你还没有安装 JDK。这时，你可以去 [Oracle 的官网](#) 去下载安装 JDK，然后配置好环境变量。

同样，我们需要确保 Python 也已经被安装在本地了。在命令行输入 “Python” 或者 “Python3”，如果可以成功进入交互式的 Python Shell，就说明已经安装了 Python。否则，需要去 [Python 官网](#) 下载安装 Python。这里，我推荐你使用 Python3 而不是 Python2。

我们同样可以在本地预装好 Hadoop。Spark 可以脱离 Hadoop 运行，不过有时我们也需要依赖于 HDFS 和 YARN。所以，这一步并不是必须的，你可以自行选择。

接下来我们就可以安装 Spark。首先去 [Spark 官网](#) 的下载界面。在第一个下拉菜单里选择最新的发布，第二个菜单最好选择与 Hadoop 2.7 兼容的版本。因为有时我们的 Spark 程序会依赖于 HDFS 和 YARN，所以选择最新的 Hadoop 版本比较好。

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.4.3-bin-hadoop2.7.tgz](#)
4. Verify this release using the 2.4.3 [signatures](#), [checksums](#) and [project release KEYS](#).

Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.

Link with Spark

Spark artifacts are [hosted in Maven Central](#). You can add a Maven dependency with the following coordinates:

```
groupId: org.apache.spark
artifactId: spark-core_2.11
version: 2.4.3
```

Installing with PyPi

PySpark is now available in `pypi`. To install just run `pip install pyspark`.

Release Notes for Stable Releases

- [Spark 2.4.3](#) (May 07 2019)
- [Spark 2.3.3](#) (Feb 15 2019)

Archived Releases

As new Spark releases come out for each development stream, previous ones will be archived, but they are still available at [Spark release archives](#).

Latest News

[Spark 2.4.3 released](#) (May 08, 2019)

[Spark 2.4.2 released](#) (Apr 23, 2019)

[Spark 2.4.1 released](#) (Mar 31, 2019)

[Spark 2.3.3 released](#) (Feb 15, 2019)

[Archive](#)

 **APACHECON**
LAS VEGAS: Sept. 9-12, 2019
BERLIN: Oct. 22-24, 2019

[Download Spark](#)


Built-in Libraries:

[SQL and DataFrames](#)
[Spark Streaming](#)
[MLlib \(machine learning\)](#)
[GraphX \(graph\)](#)

[Third-Party Projects](#)


Apache Spark, Spark, Apache, the Apache feather logo, and the Apache Spark project logo are either registered trademarks or trademarks of The Apache Software Foundation in the United States and other countries. See guidance on use of Apache Spark [trademarks](#). All other marks mentioned may be trademarks or registered trademarks of their respective owners. Copyright © 2018 The Apache Software Foundation. Licensed under the [Apache License, Version 2.0](#).

下载好之后，解压缩 Spark 安装包，并且把它移动到 `/usr/local` 目录下，在终端中输入下面的代码。

 复制代码

```
1 $ tar -xzf ~/Downloads/spark-2.4.3-bin-hadoop2.7.tg
2 $ mv spark-2.4.3-bin-hadoop2.7.tgz /usr/local/spark
```

经过上述步骤，从官网下载并安装 Spark 的文件，这样我们便完成了 Spark 的安装。但是，Spark 也是要进行相应的环境变量配置的。你需要打开环境变量配置文件。

 复制代码

```
1 vim ~/.bash_profile
```

并在最后添加一段代码。

 复制代码

```
1 export SPARK_HOME=/usr/local/spark
2 export PATH=$PATH:$SPARK_HOME/bin
```

这样，所需的步骤都做完之后，我们在命令行控制台输入 PySpark，查看安装情况。如果出现下面的欢迎标志，就说明安装完毕了。

 复制代码

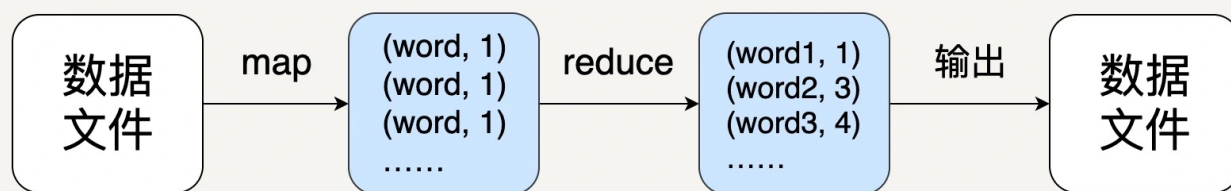
```

1 Welcome to
2
3      ____          __
4     /  _/\_/_   ___  _____/  _/_
5    _\  \/_  _  \/_  _`/_  _/_  ' _/_
6   /__  /  .___/\_,_/_/_/  _/_\_\_  version 2.4.3
7       /_/_
8
9 Using Python version 2.7.10 (default, Oct 6 2017 22:29:07)
10 SparkSession available as 'spark'.

```

基于 RDD API 的 Word Count 程序

配置好所需的开发环境之后，下一步就是写一个 Python 程序去统计词语频率。我们都知道这个程序的逻辑应该是如下图所示的。




对于中间的先 map 再 reduce 的处理，我相信通过前面的学习，所有同学都可以用 RDD 或者 DataFrame 实现。

但是，我们对于 Spark 程序的入口是什么、如何用它读取和写入文件，可能并没有了解太多。所以，接下来让我们先接触一下 Spark 程序的入口。

在 Spark 2.0 之前，**SparkContext** 是所有 Spark 任务的入口，它包含了 Spark 程序的基本设置，比如程序的名字、内存大小、并行处理的粒度等，Spark 的驱动程序需要利用它来连接到集群。


无论 Spark 集群有多少个节点做并行处理，每个程序只可以有唯一的 SparkContext，它可以被 SparkConf 对象初始化。

 复制代码

```
1 conf = SparkConf().setAppName(appName).setMaster(master)
2 sc = SparkContext(conf=conf)
```

这个 appName 参数是一个在集群 UI 上展示应用程序的名称，master 参数是一个 Spark、Mesos 或 YARN 的集群 URL，对于本地运行，它可以被指定为 “local”。

在统计词频的例子中，我们需要通过 SparkContext 对象来读取输入文件，创建一个 RDD，如下面的代码所示。

 复制代码

```
1 text_file = sc.textFile("file:///.....") //替换成实际的本地文件路径。
```

这里的 text_file 是一个 RDD，它里面的每一个数据代表原文本文件中的一行。

在这些版本中，如果要使用 Spark 提供的其他库，比如 SQL 或 Streaming，我们就需要为它们分别创建相应的 context 对象，才能调用相应的 API，比如的 DataFrame 和 DStream。

 复制代码

```
1 hc = HiveContext(sc)
2 ssc = StreamingContext(sc)
```

在 Spark 2.0 之后，随着新的 DataFrame/DataSet API 的普及化，Spark 引入了新的 **SparkSession** 对象作为所有 Spark 任务的入口。

SparkSession 不仅有 SparkContext 的所有功能，它还集成了所有 Spark 提供的 API，比如 DataFrame、Spark Streaming 和 Structured Streaming，我们再也不用为不同的功能分别定义 Context。

在统计词频的例子中，我们可以这样初始化 SparkSession 以及创建初始 RDD。

 复制代码


```
1 spark = SparkSession
2     .builder
3     .appName(appName)
4     .getOrCreate()
5 text_file = spark.read.text("file://...").rdd.map(lambda r: r[0])
```

由于 SparkSession 的普适性，我推荐你尽量使用它作为你们 Spark 程序的入口。随后的学习中，我们会逐渐了解怎样通过它调用 DataFrame 和 Streaming API。

让我们回到统计词频的例子。在创建好代表每一行文本的 RDD 之后，接下来我们便需要两个步骤。

1. 把每行的文本拆分成一个个词语；
2. 统计每个词语的频率。


对于第一步，我们可以用 flatMap 去把行转换成词语。对于第二步，我们可以先把每个词语转换成 (word, 1) 的形式，然后用 reduceByKey 去把相同词语的次数相加起来。这样，就很容易写出下面的代码了。

 复制代码

```
1 counts = lines.flatMap(lambda x: x.split(' '))
2           .map(lambda x: (x, 1))
3           .reduceByKey(add)
```

这里 counts 就是一个包含每个词语的 (word, count) pair 的 RDD。

相信你还记得，只有当碰到 action 操作后，这些转换动作才会被执行。所以，接下来我们可以用 collect 操作把结果按数组的形式返回并输出。

 复制代码

```
1 output = counts.collect()
2 for (word, count) in output:
3     print("%s: %i" % (word, count))
4 spark.stop() // 停止SparkSession
```

基于 DataFrame API 的 Word Count 程序

讲完基于 RDD API 的 Word Count 程序，接下来让我们学习下怎样用 DataFrame API 来实现相同的效果。

在 DataFrame 的世界中，我们可以把所有的词语放入一张表，表中的每一行代表一个词语，当然这个表只有一列。我们可以对这个表用一个 groupBy() 操作把所有相同的词语聚合起来，然后用 count() 来统计出每个 group 的数量。

但是问题来了，虽然 Scala 和 Java 支持对 DataFrame 进行 flatMap 操作，但是 Python 并不支持。那么要怎样把包含多个词语的句子进行分割和拆分呢？这就要用到两个新的操作——explode 和 split。split 是 pyspark.sql.functions 库提供的一个函数，它作用于 DataFrame 的某一列，可以把列中的字符串按某个分隔符分割成一个字符串数组。

explode 同样是 pyspark.sql.functions 库提供的一个函数，通俗点的翻译是“爆炸”，它也作用于 DataFrame 的某一列，可以为列中的数组或者 map 中每一个元素创建一个新的 Row。


由于之前代码中创建的 df_lines 这个 DataFrame 中，每一行只有一列，每一列都是一个包含很多词语的句子，我们可以先对这一列做 split，生成一个新的列，列中每个元素是一个词语的数组；再对这个列做 explode，可以把数组中的每个元素都生成一个新的 Row。这样，就实现了类似的 flatMap 功能。这个过程可以用下面的三个表格说明。

Lines
"I have a dog"
"He has a dog"

Lines	Word Array
"I have a dog"	["I", "have", "a", "dog"]
"He has a dog"	["He", "has", "a", "dog"]

Lines	Word Array	Word
"I have a dog"	["I", "have", "a", "dog"]	"I"
"I have a dog"	["I", "have", "a", "dog"]	"have"
"I have a dog"	["I", "have", "a", "dog"]	"a"
"I have a dog"	["I", "have", "a", "dog"]	"dog"
"He has a dog"	["He", "has", "a", "dog"]	"He"
"He has a dog"	["He", "has", "a", "dog"]	"has"
"He has a dog"	["He", "has", "a", "dog"]	"a"
"He has a dog"	["He", "has", "a", "dog"]	"dog"

接下来我们只需要对 Word 这一列做 `groupBy`，就可以统计出每个词语出现的频率，代码如下。

 复制代码

```

1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import *
3
4 if __name__ == "__main__":
5     spark = SparkSession
6         .builder
7         .appName('WordCount')
8         .getOrCreate()

```

```
9     lines = spark.read.text("sample.txt")
10    wordCounts = lines
11        .select(explode(split(lines.value, " ")))
12        .alias("word"))
13        .groupBy("word")
14        .count()
15    wordCounts.show()
16
17    spark.stop()
```

从这个例子，你可以很容易看出使用 DataSet/DataFrame API 的便利性——我们不需要创建 (word, count) 的 pair 来作为中间值，可以直接对数据做类似 SQL 的查询。

小结

通过今天的学习，我们掌握了如何从零开始创建一个简单的 Spark 的应用程序，包括如何安装 Spark、如何配置环境、Spark 程序的基本结构等等。

实践题

希望你可以自己动手操作一下，这整个过程只需要跑通一次，以后就可以脱离纸上谈兵，真正去解决实际问题。

欢迎你在留言中反馈自己动手操作的效果。

如果你跑通了，可以在留言中打个卡。如果遇到了问题，也请你在文章中留言，与我和其他同学一起讨论。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

精选留言 (26)



朱同学

2019-05-29

java万金油，什么都可以干，人好招，特别是我们这种偏远地区，scala，虽然开发效率高，但

是人少，难招，所以我们大数据团队选择了java。至于运行效率，py是最慢的，java和scala应该半斤八两吧



👍 20



科学Jia

2019-06-20

女同学看完2015年出的spark快速大数据分析这本书以后，再来看老师写的这些文字，觉得言简意赅，印象深刻，至于用什么语言倒无所谓了，主要是思路。后期希望老师能多说一些案例和处理中需要注意的技巧。



👍 9



—

2019-05-29

看了这一讲意识到之前对Python欠缺了重视，现在明白Python在大数据处理领域是很有竞争力的，因为Spark和众多的库的原因，甚至超越Java，所以现在要重新重视起来Python的学习了



👍 7



hallo128

2019-06-08

【以下代码可以运行，但对df格式的操作是借助二楼的网址去找的，具体含义也不太清楚，只是可以运行出来】

#python前运行调用包

```
from pyspark.sql import SparkSession
```

```
from pyspark.sql.functions import explode
```

```
from pyspark.sql.functions import split
```

#初始化SparkSession程序入口

```
spark = SparkSession.builder.appName("WordCount").getOrCreate()
```

#读入文档

```
ds_lines = spark.read.text("/Users/apple/code_tool/spark/WordCount/demo.md")
```

#针对df特定的计算格式

```
words = ds_lines.select(
```

```
    explode(
```

```
        split(ds_lines.value, " ")
```

```
    ).alias("word")
```

```
)  
#返回的RDD进行计数  
wordCounts = words.groupBy("word").count()  
#展示  
wordCounts.show()  
#关闭spark  
spark.stop()
```



👍 6



9527

2019-05-29

```
spark_session = SparkSession.builder.appName("PySparkShell").getOrCreate()  
ds_lines = spark_session.read.textFile("README.md")  
ds = ds_lines.flatMap(lambda x: x.split(' ')).groupBy("Value").count()  
ds.show()
```

我执行这段的时候报错了

AttributeError: 'DataFrameReader' object has no attribute 'textFile'

如果把textFile()改成text()就对了

再执行flatMap那段，也报错了

AttributeError: 'DataFrame' object has no attribute 'flatMap'

是不是API变动了，我用的是2.4.3版本单机执行的

共 1 条评论 >

👍 6



hallo128

2019-06-08

“虽然 Spark 还支持 Java 和 R，但是我个人不推荐你使用。用 Java 写程序实在有些冗长，而且速度上没有优势。”

推荐使用，还是应该详细说明对比下，不能只因为自己偏好某种工具给出建议。对于spark原生来说，速度和库同步更新更快的是Scala，如果你想随时用到spark最新功能库的话，就应该选择Scala，同时速度也是最快的。

至于Python，R，Java，一方面和你的熟悉程度有关，另一方面也与你到底准备用spark来做什么的目的有关。是集群控制，还是数据分析，还是建模，来选择合适的编程语言与spark进行连接编写。

共 1 条评论 >

👍 4



青石

2019-05-31

```
#!/usr/bin/python3

import os
from pyspark import SparkContext, SparkConf

os.environ['SPARK_HOME'] = '/usr/local/spark'
os.environ['HADOOP_HOME'] = '/usr/local/hadoop-2.7.7'

conf = SparkConf().setAppName('WordCount').setMaster('local')
sc = SparkContext('local', 'pyspark', conf=conf)

text_file = sc.textFile('file:///Users/albert.ming.xu/Downloads/text.txt')

counts = text_file.filter(lambda x: len(x.strip()) > 0).flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(lambda x, y: x + y).sortBy(lambda x: x[1], ascending=False)

print('{0: ^20}{1: ^20}'.format('Word', 'Count'))
for (word, num) in counts.take(10):
    print('{0: ^20}{1: ^20}'.format(word, num))
```



👍 3



Quincy

2019-06-13

Spark 不应该是首选Scala 么



👍 3



J Zhang

2019-05-29

用java写 有点冗长 我不敢苟同，因为java8 已经是函数编程了！而且spark开发我觉得大部分还是spark sql多点！这样基本没啥区别



👍 3



这个名字居然都有

2019-05-29

老师，你给一个完整的案例吧，

共 1 条评论 >

👍 3



斯盖丸

2019-05-29

.groupBy("Value")这个value是什么意思？

作者回复: SparkSession.read.text()读取文件后生成的DataFrame只有一列，它的默认名字就是"value"。我们用lines.value去读取这一列，是同样的道理。之后我们给新的列重命名为"word"，所以groupBy的参数变成了"word"。



👍 2



大志

2019-05-29

老师，本地已经安装了Spark，有Demo吗，只看代码片段的话还是无从下手啊

共 1 条评论 >

👍 2



JustDoDT

2019-09-01

python 直接安装

pip install pyspark

pip帮你搞定一切安装配置问题。

参考资料：

<https://pypi.org/project/pyspark/>

作者回复: 嗯，这位同学说的很好，用pip install安装pyspark确实方便。我介绍的方法比较普遍试用。

共 2 条评论 >

👍 1



Bing

2019-05-29

flatMap是rdd的算子，df不能直接用，可以explode行转列



👍 1



xxx

2021-12-28

文中的示例是可以运行的，稍微改改：


```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

if __name__ == "__main__":
    spark = SparkSession.builder.appName('WordCount').getOrCreate()
    lines = spark.read.text("wikiOfSpark.txt")
    wordCounts = lines.select(explode(split(lines.value, " ")).alias("word")).groupBy("word").
count().sort(desc("count"))
    wordCounts.show()

    spark.stop()
```



stars

2021-03-28

前面还好，到这里看不懂了，环境搭建完成，代码怎么执行，完全走不下去，是不是简单说一下



黑黑白

2021-02-09

```
from operator import add
```

```
from pyspark.sql import SparkSession
```

```
if __name__ == "__main__":
    spark = SparkSession.builder.appName("rdd")\
        .config("spark.driver.bindAddress", "127.0.0.1")\
        .getOrCreate()
    lines = spark.read.text("file:///mnt/d/playground/bigdata/spark001/sample.txt")\
        .rdd.map(lambda r: r[0])
    counts = lines.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(add)
    output = counts.collect()

    for (word, count) in output:
```

```
print("%s: %i" % (word, count))  
spark.stop()
```



whatever

2020-12-01

给所有小白到我这个程度踩了初级坑的人：

如果运行pyspark报错 ERROR SparkContext: Error initializing SparkContext，说明需要修改主机名

```
vim ~/.bash_profile
```

添加

```
export SPARK_LOCAL_HOSTNAME=localhost
```

编辑完后重新加载，执行

```
source ~/.bash_profile
```

再运行pyspark试试



姜江国

2020-03-23

在pyspark中执行的程序，为什么在spark的管控台上，看不到对应的Application呢？



刘润森

2020-03-13

我在centos搭建Spark集群，怎么用Pyspark在window连接spark集群

