1. Spark的基本使用

1.1. 执行第一个Spark程序

利用Spark自带的例子程序执行一个求PI（蒙特卡洛算法）的程序：

```
$SPARK_HOME/bin/spark-submit \
--class org.apache.spark.examples.SparkPi \
--master spark://bigdata02:7077,bigdata04:7077 \
--executor-memory 512m \
--total-executor-cores 2 \
$SPARK_HOME/examples/jars/spark-examples_2.11-2.4.6.jar \
100
```

理解：

```
执行通过/home/bigdata/apps/spark-2.4.6-bin-hadoop2.7/bin/spark-shell来运行。
如果不设置master那么默认运行在本机节点；如果设置 --master spark://bigdata02:7077,那么运行在Spark Standalone模式集群。
```

参数说明：

```
local
spark集群  standalone
yarn
```

```
--class      指定 jar 包的主类

--master     指定jar包提交的模式
详解如下：
1、local
本地模式，本地运行，可以调试（local 1个线程、local[*]不限线程、local[N] N个线程，理想情况下，N设置为你机器的CPU核数）
2、Spark
提交到Spark stanalone集群，有Master和Worker进程，会在container中jar包运行的时候自动生成
3、mesos
将jar包提交到mesos集群，运行在mesos资源管理器框架之上，由mesos负责资源管理，Spark负责任务调度和计算
4、YARN
将jar包提交到yarn集群，运行在yarn资源管理器框架之上，由yarn负责资源管理，Spark负责任务调度和计算
5、cloud
```

比如AWS的EC2，使用这个模式能很方便的访问Amazon的S3，Spark支持多种分布式存储系统，比如HDFS和S3

--deploy-mode　　指定jar的运行方式（默认是 client 模式）
详解如下：
1、client 模式
在提交端生成的JVM会一直等待所有计算过程全部完成才退出，它有两个功能，一个是提交，一个是监控jar包运行（测试环境下使用）
2、cluster 模式
在提交端生成的JVM在提交jar包后会退出，它只有一个提交功能，然后在某一个 worker上会生成一个Driver的JVM，该JVM执行监控jar包运行的功能，等待所有代码运行完毕退出（生产环境下使用 ）

application.jar　　指定你的jar包的地址
arguments　　传给main()方法的参数