链滴

提交 spark 应用

作者：bian
原文链接：https://ld246.com/article/1497706829545
来源网站：链滴

提交spark应用

提交应用

在bin目录中spark-submit 脚本用于向集群启动你的应用，通过统一的接口，可以支持所有spark 支的clusterManager，所以你没有必要为你的每个应用坐一些特殊的配置。

打包你的应用程序

如果你的代码依赖与其他的项目，为了分发所有代码到集群，需要把他们和你的应用打包在一起，所你可以打一个assembly jar 或者"uber jar" 包含你的代码和依赖。sbt和maven 都有assembly 插件在创建assembly jar 时，将hadoop和spark的依赖设置成为provided，这些依赖在运行时cluster已包含了，所以不需要用户打包进来。

##Loading Configuration from a File

spark-submit 会自动从配置文件中加载默认的属性值传递给你的应用程序。默认读取的是conf/spark default.conf

通过这种方式加载配置属性可以取消spark-submit 上的参数，列如：加入spark.master 在应用程序已经设置了那么就可以在spark-submit 取消--master 选项。**总之，通过SparkConf 设置的属性具最高的优先级，其次是通过spark-submit 传递的，最后是配置文件中默认的。**

用spark-submit 启动你的应用

一旦你的应用打包好了，你就可以用spark-submit 脚本启动你的应用了。这个脚本会设置spark的cla spath和它的依赖。并且可以支持spark所支持的所有**clusterManager** 和 **deployMode**

```
./bin/spark-submit \
  --class <main-class> \
  --master <master-url> \
  --deploy-mode <deploy-mode> \
  --conf <key>=<value> \
  ... # other options
  <application-jar> \
  [application-arguments]
```

下面讲解一些通用的选项：

- --class:程序的切入点（eg:org.apache.spark.examples.SparkPi）
- --master:集群的master url（eg.spark://xxx.xxx.xxx.xxx:xxx）
- --deploy-mode:决定你的driver程序是否在worker node（cluster模式下），或者作为一个外部cli nt（clinet 模式）**默认值是client**
- --conf 以key=value的形式设置任意的spark properties，如果value包含空格，用双引号括起来，
- <application-jar> 你的应用程序和依赖的jar的路径，URL 必须对集群是全局可见的，列如:hdfs:// 或者 a file:// 路径必须在所有的node都可见
- application-arguments 如果有额外的参数需要传递给你的main程序，请用这个选项

一个常用的提交应用的策略是通过一个网关机器(**gateway machine**),并且这个机器在物理位置上接

你的worker node，这种方式client模式是非常适合的。在client 模式下，dirver直接被spark-submi启动扮演集群一个client的角色。应用的输出都在console上，因此这种模式非常十分的适合那些需要EPL的应用（比如Spark shell）.

有一个可用选项可以指定clusterManager，在standalone集群中，cluster 模式下，你可以指定--supervise 以确保driver程序非正常退出的情况下自动重启driver。--help 显示所有可用的选项。

下面是一些可用的选项

```
# Run application locally on 8 cores
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master local[8] \
  /path/to/examples.jar \
  100

# Run on a Spark standalone cluster in client deploy mode
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master spark://207.184.161.138:7077 \
  --executor-memory 20G \
  --total-executor-cores 100 \
  /path/to/examples.jar \
  1000

# Run on a Spark standalone cluster in cluster deploy mode with supervise
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master spark://207.184.161.138:7077 \
  --deploy-mode cluster \
  --supervise \
  --executor-memory 20G \
  --total-executor-cores 100 \
  /path/to/examples.jar \
  1000

# Run on a YARN cluster
export HADOOP_CONF_DIR=XXX
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master yarn \
  --deploy-mode cluster \  # can be client for client mode
  --executor-memory 20G \
  --num-executors 50 \
  /path/to/examples.jar \
  1000

# Run a Python application on a Spark standalone cluster
./bin/spark-submit \
  --master spark://207.184.161.138:7077 \
  examples/src/main/python/pi.py \
  1000

# Run on a Mesos cluster in cluster deploy mode with supervise
```

```
./bin/spark-submit \
  --class org.apache.spark.examples.SparkPi \
  --master mesos://207.184.161.138:7077 \
  --deploy-mode cluster \
  --supervise \
  --executor-memory 20G \
  --total-executor-cores 100 \
  http://path/to/examples.jar \
  1000
```

Master URLS

master url 可以是以下的格式：

MasterURL	解释
local	本地模式一个worker线程（没有并发）
local[k]	本地模式K个worker线程（设置成你机的core数）
local[*]	本地模式多个worker线程数—>你机器逻辑core数量
spark://HOST:PORT	连接到一个spark standalone 集群，端口号必须是master指定的，7007是默认的
mesos://HOST:PORT	连接到一个Messon集群
yarn	连接到一个yarn集群。—deploy-mode 指client或者cluster模式，集群地址通过HADOOP_CONF_DIR or YARN_CONF_DIR 自动发现

通过文件配置

spark-submit 会自动从配置文件中加载默认的属性值传递给你的应用程序。默认读取的是conf/spark default.conf

通过这种方式加载配置属性可以取消spark-submit 上的参数，列如：加入spark.master 在应用程序已经设置了那么就可以在spark-submit 取消--master 选项。**总之，通过SparkConf 设置的属性具最高的优先级，其次是通过spark-submit 传递的，最后是配置文件中默认的。**