Spark Submit Applications

在Spark集群中通过bin/spark-submit脚本来提交程序，如果应用程序依赖其他项目，需要将依赖的项目jar包一起打包后再分发到Spark Cluster中，Assembly jar包括了应用程序和依赖。可以使用Maven plugin来实现程序的打包。

Spark Submit的执行命令格式如下：

*./bin/spark-submit \*

*--class <main-class> \*

*--master <master-url> \*

*--deploy-mode <deploy-mode> \*

*--conf <key>=<value> \*

*... # other options*

*<application-jar> \*

*[application-arguments]*

参数描述如下：

|  |  |
| --- | --- |
| --class | 应用程序的entry point，例如org.apache.spark.examples.SparkPi |
| --master | 集群的master URL，例如spark://23.195.26.187:7077 |
| --deploy-mode | Driver的部署方式，cluster(driver启动在worker节点上)，client，driver启动在本地客户端（默认的方式） |
| --conf | Spark的配置参数，以key=value的格式 |
| application-jar | 应用的bundled jar包，包含应用和依赖，在集群可以访问到的位置 |
| application-arguments | 传递给应用main的参数 |

如果提交主机在集群内进行Spark Submit，可以使用client mode，driver启动在客户端，输入和输出都到控制台上。这种模式适用于涉及到REPL的程序，例如Spark Shell；提交的主机在集群外，一般使用cluster mode来减少drivers和executors的网络延迟，但是在standalone模式下，Python应用并不支持cluster mode。对于Python应用，传递.py文件而不是application-jar，通过—py-file来指定.zip,.egg和.py文件。很多其他参数，例如—supervise参数，可以让程序异常情况下自动重启driver。

1. 示例1，Run application locally on 8 cores

*./bin/spark-submit \*

*--class org.apache.spark.examples.SparkPi \*

*--master local[8] \*

*/path/to/examples.jar \*

*100*

1. 示例2，Run on a Spark standalone cluster in client deploy mode

命令如下所示：

*./bin/spark-submit \*

*--class org.apache.spark.examples.SparkPi \*

*--master spark://fys1.cmss.com:7077 \*

*--executor-memory 2G \*

*--total-executor-cores 2 \*

*/path/to/examples.jar \*

*1000*

执行结果如下所示：

*17/06/25 02:46:16 INFO SparkContext: Running Spark version 2.1.1*

*17/06/25 02:46:16 WARN SparkContext: Support for Java 7 is deprecated as of Spark 2.0.0*

*17/06/25 02:46:17 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(spark); groups with view permissions: Set(); users with modify permissions: Set(spark); groups with modify permissions: Set()*

*17/06/25 02:46:17 INFO Utils: Successfully started service 'sparkDriver' on port 40963.*

*17/06/25 02:46:17 INFO SparkEnv: Registering MapOutputTracker*

*17/06/25 02:46:18 INFO Utils: Successfully started service 'SparkUI' on port 4040.*

*17/06/25 02:46:18 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://10.139.9.123:4040*

*17/06/25 02:46:18 INFO SparkContext: Added JAR file:/home/spark/spark-2.1.1-bin-hadoop2.6/examples/jars/spark-examples\_2.11-2.1.1.jar at spark://10.139.9.123:40963/jars/spark-examples\_2.11-2.1.1.jar with timestamp 1498329978222*

*17/06/25 02:46:18 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://fys1.cmss.com:7077...*

*17/06/25 02:47:38 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool*

*17/06/25 02:47:38 INFO DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:38) finished in 10.124 s*

*17/06/25 02:47:38 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 10.527359 s*

*Pi is roughly 3.141378791413788*

1. 示例3，Run on a Spark standalone cluster in cluster deploy mode with supervise

*./bin/spark-submit \*

*--class org.apache.spark.examples.SparkPi \*

*--master spark://fys1.cmss.com:6066 \*

*--deploy-mode cluster \*

*--supervise \*

*--executor-memory 20G \*

*--total-executor-cores 100 \*

*/path/to/examples.jar \*

*1000*

结果如下：

*Running Spark using the REST application submission protocol.*

*17/06/25 02:57:56 INFO RestSubmissionClient: Submitting a request for the status of submission driver-20170625025756-0002 in spark://fys1.cmss.com:6066.*

*17/06/25 02:57:57 INFO RestSubmissionClient: State of driver driver-20170625025756-0002 is now RUNNING.*

*17/06/25 02:57:57 INFO RestSubmissionClient: Driver is running on worker worker-20170624234045-10.139.9.125-33905 at 10.139.9.125:33905.*

*17/06/25 02:57:57 INFO RestSubmissionClient: Server responded with CreateSubmissionResponse:*

*{*

*"action" : "CreateSubmissionResponse",*

*"message" : "Driver successfully submitted as driver-20170625025756-0002",*

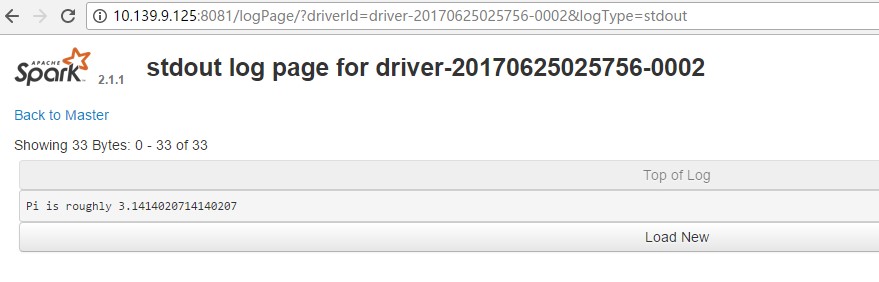
*"serverSparkVersion" : "2.1.1",*

*"submissionId" : "driver-20170625025756-0002",*

*"success" : true*

*}*

在Master Web UI中可以在日志中查看输出结果：



1. 示例4，# 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*

结果如下：

5.示例5，Spark-Submit on YARN