Spark Standalone 模式安装部署简要指南（未验证）

李传艺

1. 目标版本与环境要求

1.1 安装版本

Spark 3.0.1 (Sep 02 2020): Pre-built for Apache Hadoop 3.2 and later

下载链接：https://www.apache.org/dyn/closer.lua/spark/spark-3.0.1/spark-3.0.1-bin-hadoop3.2.tgz

1.2 环境安装：

（1）Java 8（JDK 1.8）最新版（在 8u92 版本之后的版本）

官网原话：【all you need is to have java installed on your system PATH, or the JAVA_HOME environment variable pointing to a Java installation.】

（2）Scala 2.12

参考 https://www.scala-lang.org/download/进行下载和安装。前提是已经安装 Java 8。

（3）Hadoop 3.2.1

下载地址：https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

集群部署：先安装好 Java。JDK1.8 还没有经过测试，可以先尝试使用 JDK1.8。

官网原话：【unpacking the software on all the machines in the cluster】这就是安装和部署的主要操作。

Hadoop 包括两种节点：Masters: NameNode and ResourceManager; workers: DataNode, NodeManager。在仅仅运行和使用 HDFS 时，只需要部署和启动 NameNode 和 DataNode 即可。以下为配置和启动 HDFS 的工作：

[**来源**：https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/ClusterSetup.html]

Part1：

通过 etc/hadoop/hadoop-env.sh 配置运行环境参数，至少需要配置一下 JAVA_HOME

```
export JAVA_HOME=XXX
```

其他可以选择不设置，即采用默认的配置。

官网原话：【At the very least, you must specify the JAVA_HOME so that it is correctly defined on each remote node.】

此外，可以在/etc/profile.d 中配置 HADOOP_HOME 环境变量，值为安装路径

Part2：

配置 Hadoop 进程参数，这里只需要看 HDFS 相关内容，即 NameNode 和 DataNode 内容：

etc/hadoop/core-site.xml 中的 fs.defaultFS 和 io.file.buffer.size

etc/hadoop/hdfs-site.xml 中的 dfs.namenode.name.dir, dfs.hosts, dfs.blocksize, dfs.namenode.handler.count 和 dfs.datanode.data.dir

Part3:

启动 Hadoop。前提是在集群的所有机器上都部署好了配置相同的 Hadoop。

在 NameNode 或 DataNode 上初始化 HDFS： $HADOOP_HOME/bin/hdfs namenode -format <cluster_name>

在所有 NameNode 上启动：$HADOOP_HOME/bin/hdfs --daemon start namenode

在所有 DataNode 上启动：$HADOOP_HOME/bin/hdfs --daemon start datanode

2. Standalone Cluster 部署和启动

官网原文：【To install Spark Standalone mode, you simply place a compiled version of Spark on each node on the cluster.】 只需要下载并解压到集群内的所有机器上即可

[**来源**：http://spark.apache.org/docs/latest/spark-standalone.html]

（1）在 master 上面启动 Master：./sbin/start-master.sh

启动成功后通过 http://localhost:8080 查看 master 系统信息

（2）在 worker 上启动 worker 并连接（注册）到 master 上：./sbin/start-slave.sh <master-spark-URL>

成功启动 worker 后，在 master 上刷新 localhost:8080 页面，可以看到添加的 worker 信息。

3. 测试与应用提交

安装好 Standalone Cluster 后，可以编写 spark 程序，并提交到 spark 平台执行。这里可以使用自带的 SparkPi 程序进行测试。命令如下：

```
./bin/spark-submit --class org.apache.spark.examples.SparkPi --master <master-url> /path/to/examples.jar 100
```

其中：

（1）org.apache.spark.examples.SparkPi 是任务的 main 函数所在的类

（2）master-url 是 master 的地址，例如 spark://192.168.0.1:7077

（3）/path/to/examples.jar 是这个任务的 Jar 包，所以要首先能够将自己的 spark 应用打包

（4）100 是 main 函数用到的一个参数