1. Spark基本原理

参见课程PPT

2. Spark安装配置（Spark on YARN模式）

安装主机：

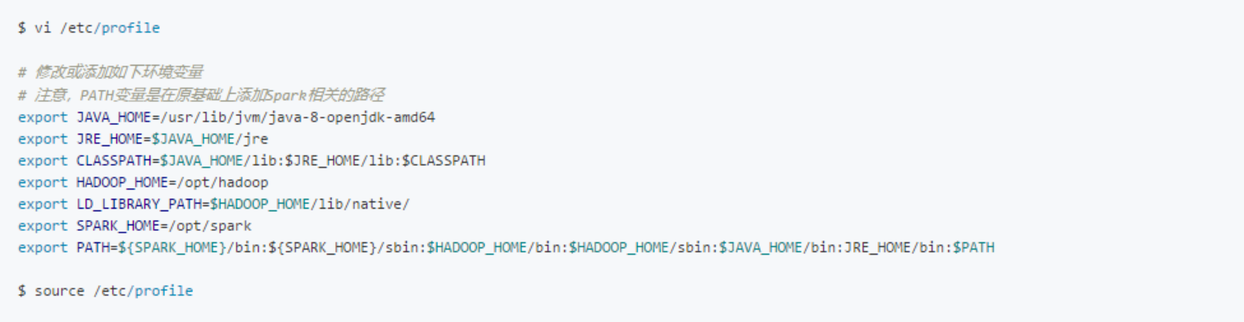
- bdcourse-0001

注意：Spark on YARN运行模式，只需要在Hadoop分布式集群中任选一个节点安装配置Spark即可，不要集群安装。因为Spark应用程序提交到YARN后，YARN会负责集群资源的调度。

安装

```
$ wget https://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-without-hadoop.tgz
$ tar -zxvf spark-2.4.5-bin-without-hadoop.tgz
$ mv spark-2.4.5-bin-without-hadoop /opt/
$ ln -s /opt/spark-2.4.5-bin-without-hadoop /opt/spark
```

配置环境变量

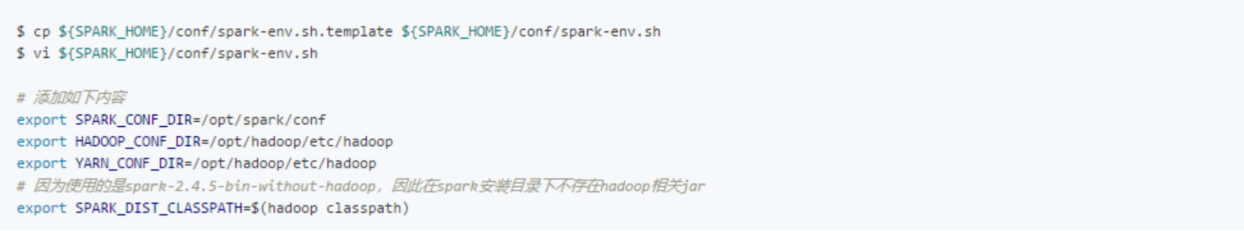


```
$ vi /etc/profile

# 修改或添加如下环境变量
# 注意，PATH变量是在原基础上添加spark相关的路径
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export JRE_HOME=$JAVA_HOME/jre
export CLASSPATH=$JAVA_HOME/lib:$JRE_HOME/lib:$CLASSPATH
export HADOOP_HOME=/opt/hadoop
export LD_LIBRARY_PATH=$HADOOP_HOME/lib/native/
export SPARK_HOME=/opt/spark
export PATH=${SPARK_HOME}/bin:${SPARK_HOME}/sbin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$JAVA_HOME/bin:JRE_HOME/bin:$PATH

$ source /etc/profile
```

配置Spark on YARN模式

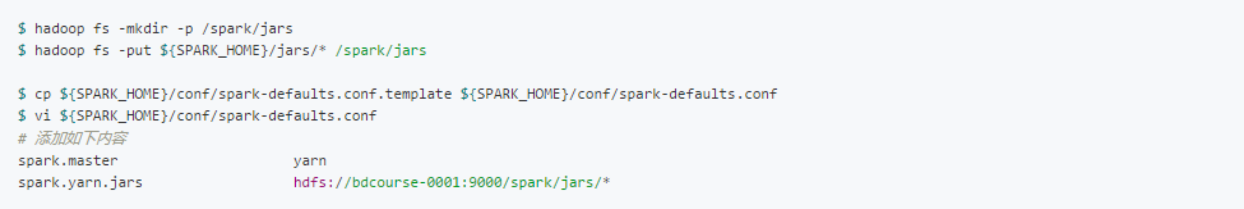


```
$ cp ${SPARK_HOME}/conf/spark-env.sh.template ${SPARK_HOME}/conf/spark-env.sh
$ vi ${SPARK_HOME}/conf/spark-env.sh

# 添加如下内容
export SPARK_CONF_DIR=/opt/spark/conf
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
export YARN_CONF_DIR=/opt/hadoop/etc/hadoop
# 因为使用的是spark-2.4.5-bin-without-hadoop，因此在spark安装目录下不存在hadoop相关jar
export SPARK_DIST_CLASSPATH=$(hadoop classpath)
```

配置公共JAR包

Hadoop 集群下提交任务时，会将 jar 包提交到 HDFS 上，为防止每次提交任务时都提交，所以在 HDFS 上上传一份公共的（将spark安装包下的jars文件夹下的jar文件全部上传上去）。并配置 `spark.yarn.jars` 为该文件夹位置（如hdfs://bdcourse-0001:9000/spark/jars/*）

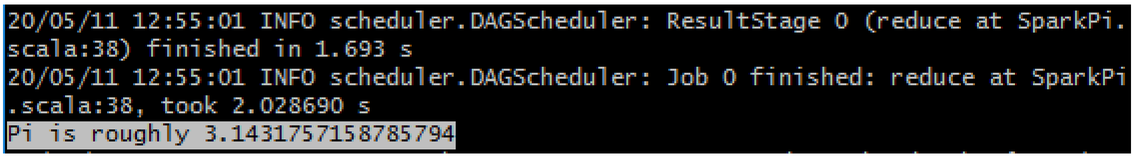


```
$ hadoop fs -mkdir -p /spark/jars
$ hadoop fs -put ${SPARK_HOME}/jars/* /spark/jars

$ cp ${SPARK_HOME}/conf/spark-defaults.conf.template ${SPARK_HOME}/conf/spark-defaults.conf
$ vi ${SPARK_HOME}/conf/spark-defaults.conf
# 添加如下内容
spark.master                     yarn
spark.yarn.jars                  hdfs://bdcourse-0001:9000/spark/jars/*
```

测试

```
# spark安装包自带scala相关jar文件，所以不需要安装scala即可运行spark-shell
$ ${SPARK_HOME}/bin/run-example SparkPi
```



```
20/05/11 12:55:01 INFO scheduler.DAGScheduler: ResultStage 0 (reduce at SparkPi.
scala:38) finished in 1.693 s
20/05/11 12:55:01 INFO scheduler.DAGScheduler: Job 0 finished: reduce at SparkPi
.scala:38, took 2.028690 s
Pi is roughly 3.1431757158785794
```

3. Zeppelin安装

Apache Zeppelin: Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.

安装主机：

- bdcourse-0001

3.1 安装

```
$ wget https://archive.apache.org/dist/zeppelin/zeppelin-0.8.2/zeppelin-0.8.2-bin-all.tgz
$ tar -xzvf zeppelin-0.8.2-bin-all.tgz
$ mv zeppelin-0.8.2-bin-all /opt/
$ ln -s /opt/zeppelin-0.8.2-bin-all /opt/zeppelin
```

3.2 配置环境变量

```
$ cp /opt/zeppelin/conf/zeppelin-env.sh.template /opt/zeppelin/conf/zeppelin-env.sh
$ chmod u+x /opt/zeppelin/conf/zeppelin-env.sh
$ vi /opt/zeppelin/conf/zeppelin-env.sh
# 添加如下内容
export SPARK_HOME=/opt/spark
export HADOOP_CONF_DIR=/opt/hadoop/etc/hadoop
export PYTHONPATH=$SPARK_HOME/python/:$PYTHONPATH
export PYTHONPATH=$SPARK_HOME/python/lib/py4j-0.10.7-src.zip:$PYTHONPATH
export PYSPARK_PYTHON=/home/bdcourse/anaconda3/bin/python
```

3.3 Zeppelin配置

注意： `zeppelin.server.addr` 和 `zeppelin.server.allowed.origins` 配置为华为云 `bdcourse-0001` 的公网IP，并在华为云安全中打开 `zeppelin.server.port` 对应的端口。

```
$ cp /opt/zeppelin/conf/zeppelin-site.xml.template /opt/zeppelin/conf/zeppelin-site.xml
$ vi /opt/zeppelin/conf/zeppelin-site.xml

<property>
  <name>zeppelin.server.addr</name>
  <value>192.168.1.31</value>
  <description>Server binding address</description>
</property>

<property>
  <name>zeppelin.server.port</name>
  <value>8080</value>
  <description>Server port.</description>
</property>

<property>
  <name>zeppelin.notebook.dir</name>
  <value>notebook</value>
  <description>path or URI for notebook persist</description>
</property>

<property>
  <name>zeppelin.server.allowed.origins</name>
  <value>http://192.168.1.31:8080, http://bdcourse-0001:8080</value>
  <description>Allowed sources for REST and WebSocket requests (i.e. http://onehost:8080,http://otherhost.com). If you leave * you are vulnerable
</property>

<property>
  <name>zeppelin.anonymous.allowed</name>
  <value>false</value>
  <description>Anonymous user allowed by default</description>
</property>
```

认证配置

```
$ cp /opt/zeppelin/conf/shiro.ini.template /opt/zeppelin/conf/shiro.ini
$ vi /opt/zeppelin/conf/shiro.ini

# 在[users]下添加用户，注释掉其他不需要的用户
bdcourse = bdcourse2021, admin
```

3.4 问题修复

主要是Zeppelin自带的jar包与Hadoop或Spark不匹配。

修复Zeppelin中io.netty.buffer.PooledByteBufAllocator.metric()Lio/netty/buffer/PooledByteBufAllocatorMetric; java.lang.NoSuchMethodError错误

参考：https://stackoverflow.com/questions/50388919/spark-2-3-java-lang-nosuchmethoderror-io-netty-buffer-pooledbytebufallocator-me

```
$ mv ${HADOOP_HOME}/share/hadoop/common/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/common/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/common/lib/

$ mv ${HADOOP_HOME}/share/hadoop/tools/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/tools/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/tools/lib/
$ mv ${HADOOP_HOME}/share/hadoop/yarn/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/yarn/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/yarn/lib/

$ mv ${HADOOP_HOME}/share/hadoop/mapreduce/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/mapreduce/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/mapreduce/lib/

$ mv ${HADOOP_HOME}/share/hadoop/hdfs/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/hdfs/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/hdfs/lib/

$ mv ${HADOOP_HOME}/share/hadoop/kms/tomcat/webapps/kms/WEB-INF/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/kms/tomcat/webapps/kms/WEB-INF/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/kms/tomcat/webapps/kms/WEB-INF/lib/

$ mv ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/netty-3.6.2.Final.jar ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/netty-3.6.2.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-3.9.9.Final.jar ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/

$ mv ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/netty-all-4.0.23.Final.jar ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/netty-all-4.0.23.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-all-4.1.42.Final.jar ${HADOOP_HOME}/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/

$ mv ${HADOOP_HOME}/share/hadoop/hdfs/lib/netty-all-4.0.23.Final.jar ${HADOOP_HOME}/share/hadoop/hdfs/lib/netty-all-4.0.23.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-all-4.1.42.Final.jar ${HADOOP_HOME}/share/hadoop/hdfs/lib/

$ mv /opt/zeppelin/lib/netty-all-4.0.23.Final.jar /opt/zeppelin/lib/netty-all-4.0.23.Final.jar.bak
$ cp ${SPARK_HOME}/jars/netty-all-4.1.42.Final.jar /opt/zeppelin/lib/

# 重启HDFS, YARN, Zeppelin
$ ${HADOOP_HOME}/sbin/stop-dfs.sh
$ ${HADOOP_HOME}/sbin/start-dfs.sh
$ ${HADOOP_HOME}/sbin/stop-yarn.sh
$ ${HADOOP_HOME}/sbin/start-yarn.sh
$ ${HADOOP_HOME}/sbin/mr-jobhistory-daemon.sh stop historyserver
$ ${HADOOP_HOME}/sbin/mr-jobhistory-daemon.sh start historyserver
```

修复Zeppelin中py4j版本过低导致使用%python手动创建spark环境后无法运行。

原因是zeppelin中py4j太旧，将spark目录下py4j复制到zeppelin中。

```
$ cp /opt/spark/jars/py4j-0.10.7.jar /opt/zeppelin/interpreter/python
# 删除原有py4j
$ mv /opt/zeppelin/interpreter/python/py4j-0.9.2.jar /opt/zeppelin/interpreter/python/py4j-0.9.2.jar.bak
$ mv /opt/zeppelin/interpreter/python/py4j-0.9.2.zip /opt/zeppelin/interpreter/python/py4j-0.9.2.zip.bak
$ mv /opt/zeppelin/interpreter/python/py4j-0.9.2 /opt/zeppelin/interpreter/python/py4j-0.9.2.bak
```

修复zeppelin spark interpreter异常 com.fasterxml.jackson.databind.JsonMappingException

zeppelin中相关jackson包太新，将zeppelin下的jackson相关包使用spark下的替换

```
$ mv /opt/zeppelin/lib/jackson-annotations-2.8.0.jar /opt/zeppelin/lib/jackson-annotations-2.8.0.jar.bak
$ mv /opt/zeppelin/lib/jackson-core-2.8.10.jar /opt/zeppelin/lib/jackson-core-2.8.10.jar.bak
$ mv /opt/zeppelin/lib/jackson-databind-2.8.11.1.jar /opt/zeppelin/lib/jackson-databind-2.8.11.1.jar.bak
$ mv /opt/zeppelin/lib/jackson-module-jaxb-annotations-2.8.10.jar /opt/zeppelin/lib/jackson-module-jaxb-annotations-2.8.10.jar.bak
$ cp /opt/spark/jars/jackson-annotations-2.6.7.jar /opt/zeppelin/lib/
$ cp /opt/spark/jars/jackson-core-2.6.7.jar /opt/zeppelin/lib/
$ cp /opt/spark/jars/jackson-databind-2.6.7.3.jar /opt/zeppelin/lib/
$ cp /opt/spark/jars/jackson-module-jaxb-annotations-2.6.7.jar /opt/zeppelin/lib/
```

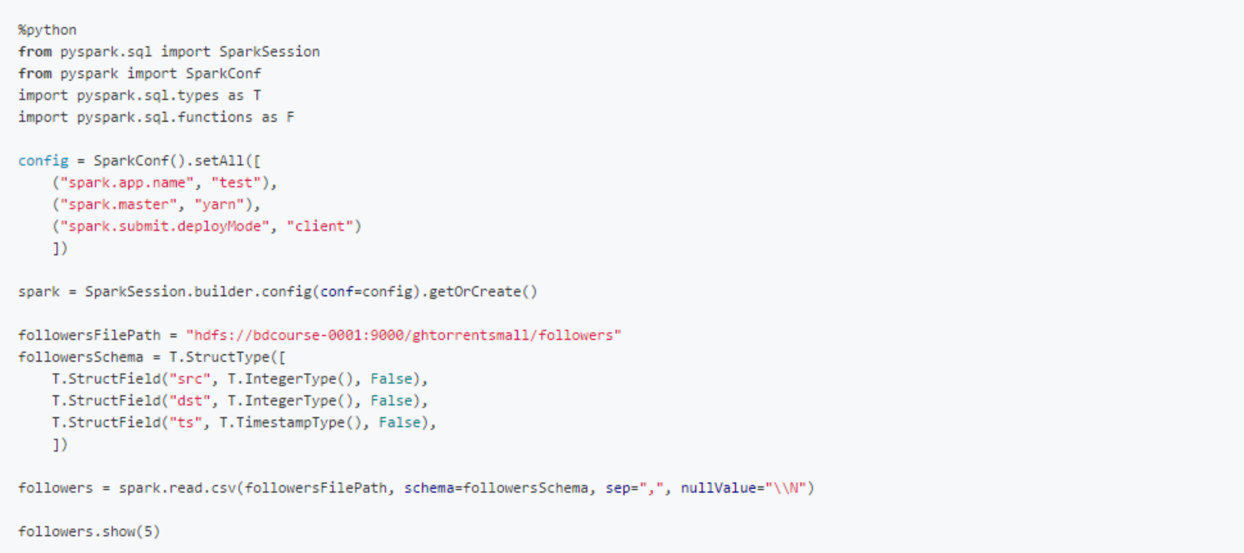
3.5 启动Zeppelin

```
$ /opt/zeppelin/bin/zeppelin-daemon.sh start
```

修复%file提示jersey出错问题

```
# 在Zeppelin Web页面Interpreter中配置file，添加artifact
/opt/hadoop/share/hadoop/common/lib/jersey-core-1.9.jar
/opt/hadoop/share/hadoop/common/lib/jersey-json-1.9.jar
/opt/hadoop/share/hadoop/common/lib/jersey-server-1.9.jar
/opt/hadoop/share/hadoop/yarn/lib/jersey-client-1.9.jar
/opt/hadoop/share/hadoop/yarn/lib/jersey-guice-1.9.jar
# hdfs.url配置为http://bdcourse-0001:50070/webhdfs/v1/
# hdfs.user 配置为  bdcourse
# %file 目前所知支持cd, ls, pwd等命令
```

3.6 Zeppelin使用示例



```
%python
from pyspark.sql import SparkSession
from pyspark import SparkConf
import pyspark.sql.types as T
import pyspark.sql.functions as F

config = SparkConf().setAll([
    ("spark.app.name", "test"),
    ("spark.master", "yarn"),
    ("spark.submit.deployMode", "client")
    ])

spark = SparkSession.builder.config(conf=config).getOrCreate()

followersFilePath = "hdfs://bdcourse-0001:9000/ghtorrentsmall/followers"
followersSchema = T.StructType([
    T.StructField("src", T.IntegerType(), False),
    T.StructField("dst", T.IntegerType(), False),
    T.StructField("ts", T.TimestampType(), False),
    ])

followers = spark.read.csv(followersFilePath, schema=followersSchema, sep=",", nullValue="\\N")

followers.show(5)
```

```
%python
followers.createOrReplaceTempView("followers")
spark.sql("SELECT * FROM followers LIMIT 3").show(3)
spark.stop()
```

4. 基于Spark MLlib的开源软件项目流行度预测

见PPT