实时数据流处理

采用框架：kafka+spark streaming

---

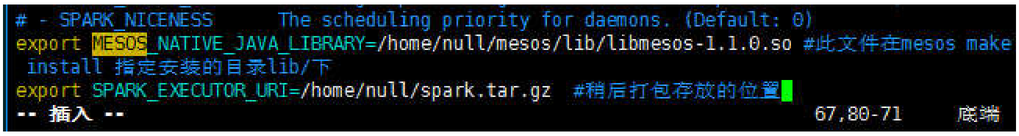
1. 环境搭建
    - kafka 环境搭建
    - spark streaming 环境搭建（Mesos）
        - mesos环境搭建，参考文档Mesos.md
        - mesos上部署spark
            1. 下载spark

               ```
               wget http://d3kbcqa49mib13.cloudfront.net/spark-2.0.2-bin-hadoop2.7.tgz
               tar zvxf spark-2.0.2-bin-hadoop2.7.tgz
               ```

            2. 配置

               ```
               cd spark-2.0.2-bin-hadoop2.7/conf
               cat spark-env.sh.template spark-env.sh
               vim spark-env.sh #参照示例，添加内容
               ```

               无hdfs版:

               ```
               # - SPARK_NICENESS      The scheduling priority for daemons. (Default: 0)
               export MESOS_NATIVE_JAVA_LIBRARY=/home/null/mesos/lib/libmesos-1.1.0.so #此文件在mesos make
                install 指定安装的目录lib/下
               export SPARK_EXECUTOR_URI=/home/null/spark.tar.gz  #稍后打包存放的位置
               -- 插入 --                                                67,80-71      底端
               ```

               hdfs版：
               修改上述示例的最后一行，即export SPARK_EXECUTOR_URI=hdfs://spark.tar.gz

               ```
               cat spark-defaults.conf.template spark-defaults.conf
               vim spark-defaults.conf #添加如下内容
               ```

               spark.io.compression.codec lzf

            3. 打包分发

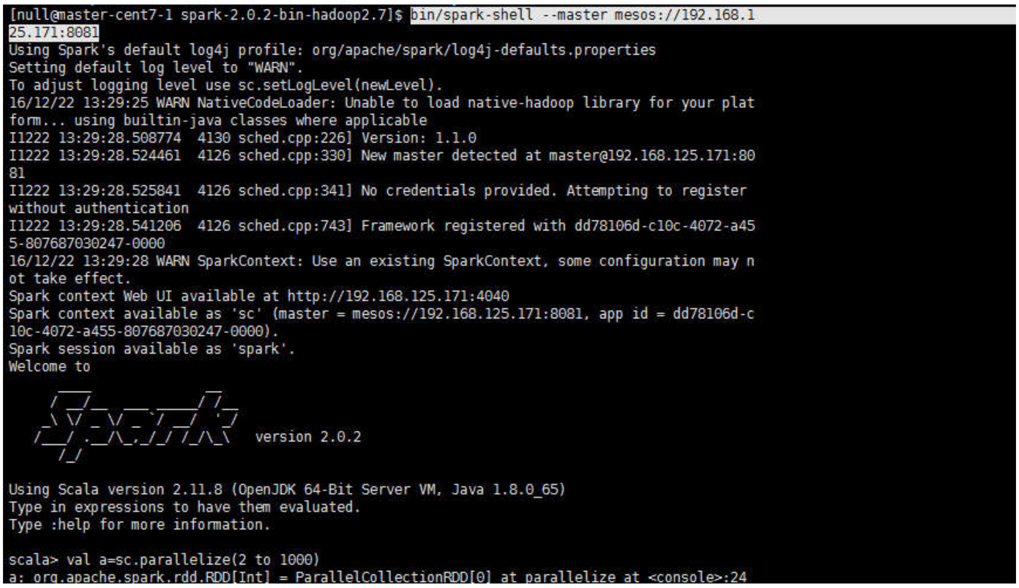
               ```
               cd ../..
               tar zcvf spark.tar.gz spark-2.0.2-bin-hadoop2.7 #压缩文件名和位置与第2步骤SPARK_EXECUTOR_URI一致
               #无hdfs
               scp spark.tar.gz 192.168.125.172:/home/null
               scp spark.tar.gz 192.168.125.173:/home/null
               ```

```
#hdfs 版，将spark.tar.gz上传到hdfs
```

4. 在master主机运行测试：

   ```
   cd spark-2.0.2-bin-hadoop2.7
   bin/spark-shell --master mesos://192.168.125.171:8081
   ```

   效果如下：

```
[null@master-cent7-1 spark-2.0.2-bin-hadoop2.7]$ bin/spark-shell --master mesos://192.168.1
25.171:8081
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/12/22 13:29:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your plat
form... using builtin-java classes where applicable
I1222 13:29:28.508774  4130 sched.cpp:226] Version: 1.1.0
I1222 13:29:28.524461  4126 sched.cpp:330] New master detected at master@192.168.125.171:80
81
I1222 13:29:28.525841  4126 sched.cpp:341] No credentials provided. Attempting to register
without authentication
I1222 13:29:28.541206  4126 sched.cpp:743] Framework registered with dd78106d-c10c-4072-a45
5-807687030247-0000
16/12/22 13:29:28 WARN SparkContext: Use an existing SparkContext, some configuration may n
ot take effect.
Spark context Web UI available at http://192.168.125.171:4040
Spark context available as 'sc' (master = mesos://192.168.125.171:8081, app id = dd78106d-c
10c-4072-a455-807687030247-0000).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.0.2
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_65)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val a=sc.parallelize(2 to 1000)
a: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:24
```

2. 上传程序

   ```
   ./sbin/start-mesos-dispatcher.sh mesos://192.168.125.171:5050
   ./bin/spark-submit \
     --class <main-class> \
     --master <master-url> \
     --deploy-mode <deploy-mode> \
     --conf <key>=<value> \
     ... # other options
     <application-jar> \
     [application-arguments]
   ```

   - 客户端模式
     上面选项中的–deploy-mode 指定为client
   - 集群模式
     上面选项中的–deploy-mode 指定为cluster
     例子：

     ```
     ./bin/spark-submit --class *** --master mesos://192.168.125.171:7077 --d
     eploy-mode client ***.jar 1000
     ```

     注意，master地址是MesosClusterDispatcher地址，默认是7077

3.