**注意事项**

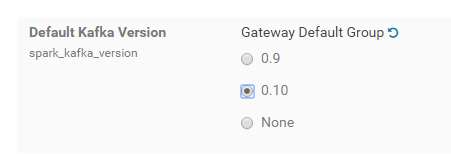
* kafka 和 spark Streaming 整合时，提交到集群运行时报错

Exception in thread "streaming-start" java.lang.NoSuchMethodError:

org.apache.kafka.clients.consumer.KafkaConsumer.subscribe(Ljava/util/Collection;)V

原因，使用kafka新型API(0.10版本)需要在CDH中进行配置

<https://www.cloudera.com/documentation/spark2/latest/topics/spark2_kafka.html#running_jobs>



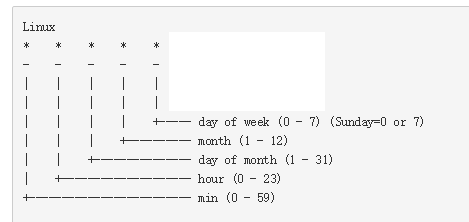
* 使用flume-ng 命令时—name后面指定的是在配置文件中配置的agent的名称
* **linux中的定时调度命令 crontab**

校验工具 https://tool.lu/crontab/

crontab -e : 选择编辑器，进入编辑文件，加入需要定时调度的语句

下面是cron的参数说明

#m h dom mon dow command



例子：

\* \* \* \* \* date >> /root/bigdata/log.txt

每分钟定时把日期写入log.txt文件

星号 表示任意值，比如在小时部分填写\*代表任意小时（每小时）

逗号 可以允许在一个部分中填写多个值，比如在分钟部分填写1,3表示一分钟或三分钟

斜线 一般配合 \*使用，代表每隔多长时间，比如在小时部分填写\*/2代表每隔两分钟。所以 \*/1和 \*没有区别

\* \* \* \* \*    # 每隔一分钟执行一次任务

0 \* \* \* \*    # 每小时的0点执行一次任务，比如6:00，10:00

6,10 \* 2 \* \*  # 每个月2号，每小时的6分和10分执行一次任务

\*/3,\*/5 \* \* \* \*  # 每隔3分钟或5分钟执行一次任务

3 \* \* \* \* 每小时的03分执行

5 3 \* \* \* 就是每天的03:05执行

\*/5 9-17 \* \* 在所有的工作日的(9-17点)每过5分钟执行

* flume与log4j进行整合

使用lo4j Appender

log4j作为flume agent的avro source

log4j.xml中的配置

*#...*

log4j.appender.flume **=** org.apache.flume.clients.log4jappender.Log4jAppender

log4j.appender.flume.Hostname **=** example.com

log4j.appender.flume.Port **=** 41414

log4j.appender.flume.UnsafeMode **=** true

*# configure a class's logger to output to the flume appender*

log4j.logger.org.example.MyClass **=** DEBUG,flume

*#...*

需要导入 flume-ng-sdk-1.8.0.jar包

在代码中，使用logger.log(msg)产生日志，自动发送到flume的source中。

