spark_streaming_hands_on

This note is written by Rudy Lee

環境

Vmware Workstaion - Centos7(VMware Workstation 15 Pro)

kafka

spark - 2.4.6

架設配置(以M、s1、s2代稱)

- tibame@master(Memory 16g、Processors 2、HDD 100g)
- tibame@slave1(Memory 5g、Processors 1、HDD 20g)
- tibame@slave2(Memory 5g、Processors 1、HDD 20g)

kafka參照Install Kafka.txt設定

啟動zookeeper

```
[tibame@master kafka_2.12-2.6.0]$ bin/zookeeper-server-start.sh config/zookeeper.properties
[2020-09-20 18:49:24,906] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2020-09-20 18:49:24,907] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2020-09-20 18:49:24,911] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2020-09-20 18:49:24,911] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2020-09-20 18:49:24,913] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2020-09-20 18:49:24,913] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2020-09-20 18:49:24,913] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2020-09-20 18:49:24,913] WARN Either no config or no quorum defined in config, running  in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2020-09-20 18:49:24,936] INFO Log4j 1.2 jmx support found and enabled. (org.apache.zookeeper.jmx.ManagedUtil)
```

分別啟動kafka三台server

```
[tibame@master kafka_2.12-2.6.0]$ bin/kafka-server-start.sh config/server-0.properties
[2020-09-20 18:49:33,823] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2020-09-20 18:49:34,709] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2020-09-20 18:49:34,781] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2020-09-20 18:49:34,786] INFO starting (kafka.server.KafkaServer)
[2020-09-20 18:49:34,787] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
```

```
[tibame@master kafka_2.12-2.6.0]$ bin/kafka-server-start.sh config/server-1.properties
[2020-09-17 23:01:21,697] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2020-09-17 23:01:22,316] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2020-09-17 23:01:22,396] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2020-09-17 23:01:22,403] INFO starting (kafka.server.KafkaServer)
[2020-09-17 23:01:22,404] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
```

```
[tibame@master kafka_2.12-2.6.0]$ bin/kafka-server-start.sh config/server-2.properties
[2020-09-17 23:01:47,982] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2020-09-17 23:01:48,718] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2020-09-17 23:01:48,763] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2020-09-17 23:01:48,782] INFO starting (kafka.server.KafkaServer)
[2020-09-17 23:01:48,783] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
```

檢查jps執行緒(確認)

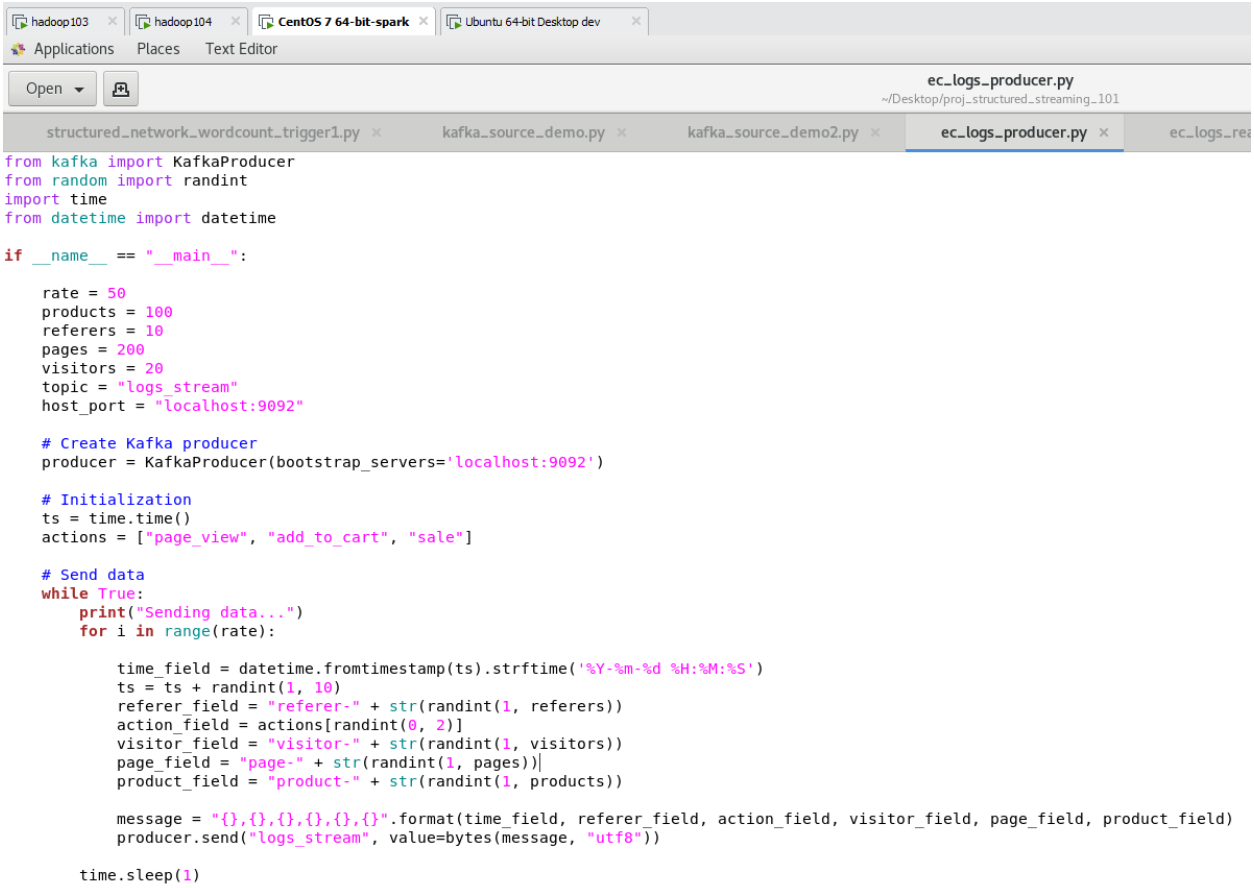
```
[tibame@master kafka_2.12-2.6.0]$ jps
79105 ResourceManager
58656 SparkSubmit
64160 SparkSubmit
73890 Master
78631 DataNode
85223 SparkSubmit
27718 SparkSubmit
3690 SparkSubmit
120682 QuorumPeerMain
18924 SparkSubmit
124337 Kafka
79251 NodeManager
55956 SparkSubmit
78423 NameNode
78873 SecondaryNameNode
19291 SparkSubmit
49631 SparkSubmit
101790 Jps
```

確認kafka已啟動

檢查zookeeper執行

```
[tibame@master kafka_2.12-2.6.0]$ sudo lsof -i :2181
COMMAND    PID   USER   FD   TYPE    DEVICE SIZE/OFF NODE NAME
java    105434 tibame  116u  IPv6 157673805      0t0  TCP *:eforward (LISTEN)
java    105434 tibame  120u  IPv6 157678050      0t0  TCP localhost:eforward->localhost:55954 (ESTABLISHED)
java    105434 tibame  122u  IPv6 157678470      0t0  TCP localhost:eforward->localhost:55956 (ESTABLISHED)
java    105434 tibame  123u  IPv6 157679948      0t0  TCP localhost:eforward->localhost:55962 (ESTABLISHED)
java    106074 tibame  116u  IPv6 157678049      0t0  TCP localhost:55954->localhost:eforward (ESTABLISHED)
java    106486 tibame  116u  IPv6 157678469      0t0  TCP localhost:55956->localhost:eforward (ESTABLISHED)
java    106868 tibame  116u  IPv6 157679947      0t0  TCP localhost:55962->localhost:eforward (ESTABLISHED)
```

測試資料(產生produce資料ec_logs_producer.py)



Terminal1(執行produce資料)

```
[tibame@master proj_structured_streaming_101]$ spark-submit --master spark://master:7077 --packages org.apache.spark:spark-streaming-kafka-0-8_2.11:2.4.5 --total-executor-cores 1 ec_logs_producer.py
```

Terminal2(確認consume到資料)

```
[tibame@master kafka_2.12-2.6.0]$ bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic logs_stream --from-beginning
```

執行結果