kafka分区重分布功能

以reassign-test这个topic作为测试

1.查看分区的分布情况

reassign-test的分区副本分布在broker1001，1003，1005，1006上

```
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --
zookeeper 192.168.2.27:12181/kafka-auth-test-1 --topic reassign-test
```

执行命令

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --zookeeper 192.168.2.27:12181/kafka
-auth-test-1 --topic reassign-test
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Topic:reassign-test     PartitionCount:3        ReplicationFactor:3     Configs:
        Topic: reassign-test    Partition: 0    Leader: 1001    Replicas: 1001,1003,1005        Isr: 1001,1003,1005
        Topic: reassign-test    Partition: 1    Leader: 1003    Replicas: 1003,1005,1006        Isr: 1003,1005,1006
        Topic: reassign-test    Partition: 2    Leader: 1005    Replicas: 1005,1003,1006        Isr: 1005,1003,1006
```

2. 编辑需要重新分区的topic的json

可以指定一个或多个topic，如果分区多，数据量大的topic，copy副本会占用资源，可以通过参数–throttle指定限流速度，可以分批次进行，下面例子只选择一个topic

```
vi /data8/apps-dev/object-log/object-log-etlday/reassign.json
{
     "topics":[
                {
                        "topic":"reassign-test"
                }
        ],
     "version":1
}

--------
topic
{"topics": [{"topic": "topicA"},
            {"topic": "topicB"}],
"version":1
}
```

3. 重新指定分区分布的broker

这里的broker-list 1001,1003,1005的主机不能比副本数少

```
/usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --
zookeeper 192.168.2.27:12181/kafka-auth-test-1 --generate --topics-to-
move-json-file /data8/apps-dev/object-log/object-log-etlday/reassign.
json --broker-list 1001,1003,1005
```

执行生成分布json命令

current partition relica assignment是当前的，重新分配后的是proposed partition reassignment configuration

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper 192.168.2.27:12181/kaf
ka-auth-test-1 --generate --topics-to-move-json-file /data8/apps-dev/object-log/object-log-etlday/reassign.json --broker-list 1001,1003
,1005
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Current partition replica assignment
{"version":1,"partitions":[{"topic":"reassign-test","partition":2,"replicas":[1005,1003,1006],"log_dirs":["any","any","any"]},{"topic":
"reassign-test","partition":1,"replicas":[1003,1005,1006],"log_dirs":["any","any","any"]},{"topic":"reassign-test","partition":0,"repli
cas":[1001,1003,1005],"log_dirs":["any","any","any"]}]}

Proposed partition reassignment configuration
{"version":1,"partitions":[{"topic":"reassign-test","partition":0,"replicas":[1001,1003,1005],"log_dirs":["any","any","any"]},{"topic":
"reassign-test","partition":2,"replicas":[1005,1003,1001],"log_dirs":["any","any","any"]},{"topic":"reassign-test","partition":1,"repli
cas":[1003,1005,1001],"log_dirs":["any","any","any"]}]}
```

4. 将生成的json保存为一个文件

```
vi /data8/apps-dev/object-log/object-log-etlday/project_new.json

{"version":1,"partitions":[{"topic":"reassign-test","partition":0,"
replicas":[1001,1003,1005],"log_dirs":["any","any","any"]},{"topic":"
reassign-test","partition":2,"replicas":[1005,1003,1001],"log_dirs":
["any","any","any"]},{"topic":"reassign-test","partition":1,"replicas":
[1003,1005,1001],"log_dirs":["any","any","any"]}]}
```

5.运行命令进行分区重新分配

```
/usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --
zookeeper ctl-nm-hhht-yxxya6-ceph-027.ctyuncdn.net:12181/kafka-auth-
test-1 --execute --reassignment-json-file /data8/apps-dev/object-log
/object-log-etlday/project_new.json
```

执行命令

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper ctl-nm-hhht-yxxya6-cep
h-027.ctyuncdn.net:12181/kafka-auth-test-1 --execute --reassignment-json-file /data8/apps-dev/object-log/object-log-etlday/project_new.
json
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Current partition replica assignment

{"version":1,"partitions":[{"topic":"reassign-test","partition":2,"replicas":[1005,1003,1006],"log_dirs":["any","any","any"]},{"topic":
"reassign-test","partition":1,"replicas":[1003,1005,1006],"log_dirs":["any","any","any"]},{"topic":"reassign-test","partition":0,"repli
cas":[1001,1003,1005],"log_dirs":["any","any","any"]}]}

Save this to use as the --reassignment-json-file option during rollback
Successfully started reassignment of partitions.
```

查看重新分配进度

```
/usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --
zookeeper ctl-nm-hhht-yxxya6-ceph-027.ctyuncdn.net:12181/kafka-auth-
test-1 --verify --reassignment-json-file /data8/apps-dev/object-log
/object-log-etlday/project_new.json
```

执行命令

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper ctl-nm-hhht-yxxya6-cep
h-027.ctyuncdn.net:12181/kafka-auth-test-1 --verify --reassignment-json-file /data8/apps-dev/object-log/object-log-etlday/project_new.j
son
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Status of partition reassignment:
Reassignment of partition reassign-test-0 completed successfully
Reassignment of partition reassign-test-2 completed successfully
Reassignment of partition reassign-test-1 completed successfully
```

6.验证

```
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --
zookeeper 192.168.2.27:12181/kafka-auth-test-1 --topic reassign-test
```

执行查看命令，副本已经没有在1006上

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --zookeeper 192.168.2.27:12181/kafka
-auth-test-1 --topic reassign-test
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Topic:reassign-test     PartitionCount:3        ReplicationFactor:3     Configs:
        Topic: reassign-test    Partition: 0    Leader: 1001    Replicas: 1001,1003,1005        Isr: 1001,1003,1005
        Topic: reassign-test    Partition: 1    Leader: 1003    Replicas: 1003,1005,1001        Isr: 1003,1005,1001
        Topic: reassign-test    Partition: 2    Leader: 1005    Replicas: 1005,1003,1001        Isr: 1005,1003,1001
```

同样的方法把g1-test在1002（已经下线）上的副本重新分配

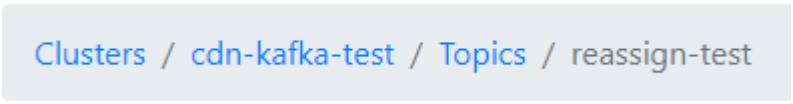
```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --zookeeper 192.168.2.27:12181/kafka
-auth-test-1 --topic g1-test
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Topic:g1-test   PartitionCount:3        ReplicationFactor:3     Configs:
        Topic: g1-test  Partition: 0    Leader: 1003    Replicas: 1003,1001,1002        Isr: 1003,1001
        Topic: g1-test  Partition: 1    Leader: 1001    Replicas: 1001,1002,1003        Isr: 1001,1003
        Topic: g1-test  Partition: 2    Leader: 1003    Replicas: 1002,1003,1001        Isr: 1003,1001
```

执行重分布命令后

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --describe --zookeeper 192.168.2.27:12181/kafka
-auth-test-1 --topic g1-test
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Topic:g1-test   PartitionCount:3        ReplicationFactor:3     Configs:
        Topic: g1-test  Partition: 0    Leader: 1003    Replicas: 1003,1007,1005        Isr: 1003,1007,1005
        Topic: g1-test  Partition: 1    Leader: 1007    Replicas: 1007,1003,1005        Isr: 1003,1007,1005
        Topic: g1-test  Partition: 2    Leader: 1003    Replicas: 1005,1003,1007        Isr: 1003,1005,1007
```

7.通过kafka manager用界面执行

从管理页面找到对应topic



Operations

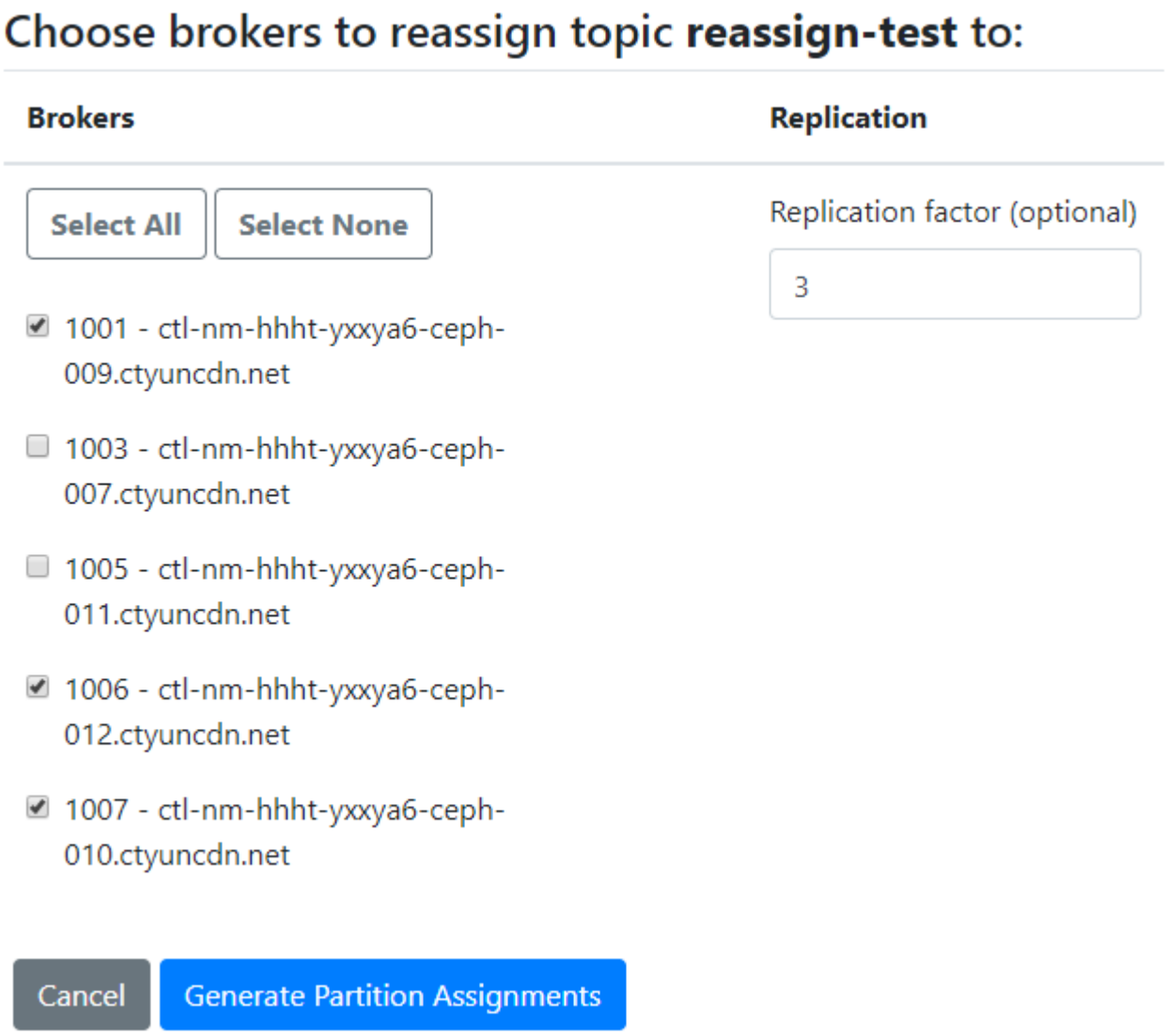
Delete Topic | Reassign Partitions | Generate Partition Assignments

Add Partitions | Update Config | Manual Partition Assignments

Partitions by Broker

Broker	# of Partitions	# as Leader	Partitions	Skewed?	Leader Skewed?
1001	3	1	(0,1,2)	false	false
1003	3	1	(0,1,2)	false	false
1005	3	1	(0,1,2)	false	false

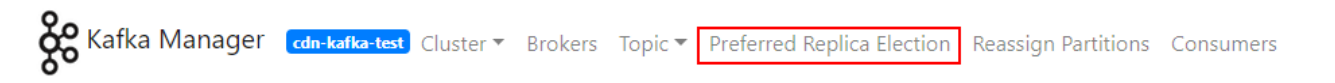
生成分区分配规则,可以通过Genrate Partition Assignment和Manual Partition Assignment两种方式

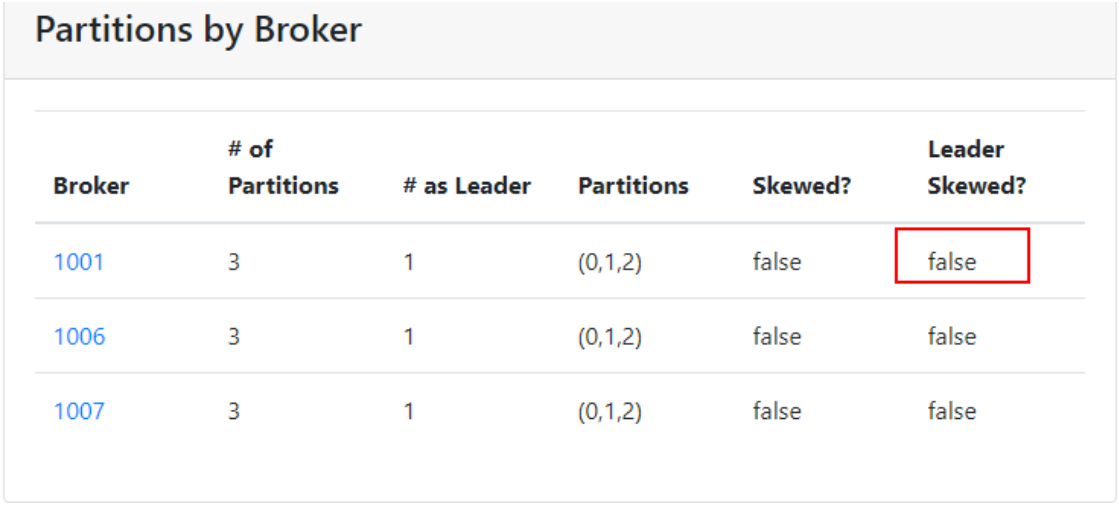


点击执行Reassign partitions，执行成功，但是第一副本不是leader副本，即1001有两个leader副本，1006没有，通过下面的步骤可以重新选举

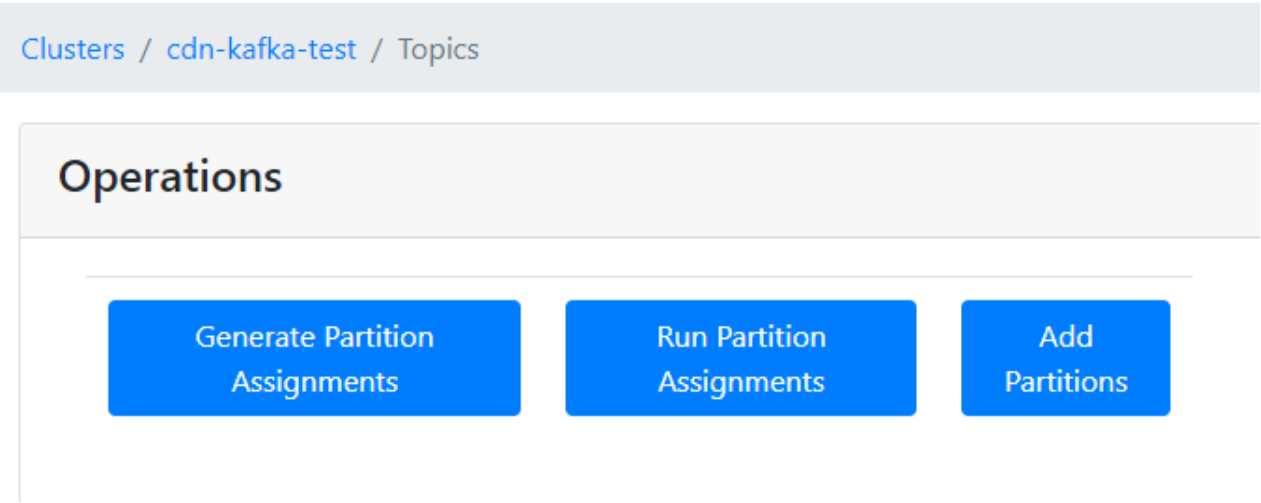
Broker	# of Partitions	# as Leader	Partitions	Skewed?	Leader Skewed?
1001	3	2	(0,1,2)	false	true
1006	3	0	(0,1,2)	false	false
1007	3	1	(0,1,2)	false	false

通过Preferred Replica Election可以重新选举优先副本





批量执行可以使用topics下面的Generate Partition Assignments和Run Partition Assignments



8.遇到的问题

执行execute，有一个存在的分配正在运行，删除zookeeper的节点rmr /admin/reassign_partitions后可以执行

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper ctl-nm-hhht-yxxya6-ceph-027.ctyuncdn.net:12181/kafka-auth-test-1 --execute --reassignment-json-file /data8/apps-dev/object-log/object-log-etlday/project_new.json
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
There is an existing assignment running.
```

检查进度，一直是still in progress，等了很久也是这个状态

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper ctl-nm-hhht-yxxya6-ceph-027.ctyuncdn.net:12181/kafka-auth-test-1 --verify --reassignment-json-file /data8/apps-dev/object-log/object-log-etlday/project_new.json
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Status of partition reassignment:
Reassignment of partition reassign-test-0 is still in progress
Reassignment of partition reassign-test-2 is still in progress
Reassignment of partition reassign-test-1 is still in progress
```

发现Controller的所在的broker是1007，查看日志，重新执行分配，日志里没有反应，考虑到可能是Controller出了问题，删掉zookeeper节点rmr /controller，会重新选举controller

```
[2020-02-20 15:55:28,586][62] DEBUG [Controller id=1007] Resigning (kafka.controller.KafkaController)
[2020-02-20 15:55:28,587][62] DEBUG [Controller id=1007] Unregister BrokerModifications handler for Set(1001, 1007, 1006, 1003, 1005) (
kafka.controller.KafkaController)
[2020-02-20 15:55:28,589][66] INFO [PartitionStateMachine controllerId=1007] Stopped partition state machine (kafka.controller.Partitio
nStateMachine)
[2020-02-20 15:55:28,590][66] INFO [ReplicaStateMachine controllerId=1007] Stopped replica state machine (kafka.controller.ReplicaState
Machine)
[2020-02-20 15:55:28,594][66] INFO [RequestSendThread controllerId=1007] Shutting down (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,594][66] INFO [RequestSendThread controllerId=1007] Stopped (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,594][66] INFO [RequestSendThread controllerId=1007] Shutdown completed (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,596][66] INFO [RequestSendThread controllerId=1007] Shutting down (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,597][66] INFO [RequestSendThread controllerId=1007] Stopped (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,597][66] INFO [RequestSendThread controllerId=1007] Shutdown completed (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,598][66] INFO [RequestSendThread controllerId=1007] Shutting down (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,598][66] INFO [RequestSendThread controllerId=1007] Stopped (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,598][66] INFO [RequestSendThread controllerId=1007] Shutdown completed (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,600][66] INFO [RequestSendThread controllerId=1007] Shutting down (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,601][66] INFO [RequestSendThread controllerId=1007] Stopped (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,601][66] INFO [RequestSendThread controllerId=1007] Shutdown completed (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,602][66] INFO [RequestSendThread controllerId=1007] Shutting down (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,602][66] INFO [RequestSendThread controllerId=1007] Stopped (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,602][66] INFO [RequestSendThread controllerId=1007] Shutdown completed (kafka.controller.RequestSendThread)
[2020-02-20 15:55:28,604][66] INFO [Controller id=1007] Resigned (kafka.controller.KafkaController)
[2020-02-20 15:55:28,605][62] DEBUG [Controller id=1007] Broker 1005 has been elected as the controller, so stopping the election proce
ss. (kafka.controller.KafkaController)
```

发现Controller的所在的broker分配到了1005，去1005上查看Controller日志，有了reassign的记录，controller主要有下面4个功能

- Broker 的上线、下线处理；
- 新创建的 topic 或已有 topic 的分区扩容，处理分区副本的分配、leader 选举；
- 管理所有副本的状态机和分区的状态机，处理状态机的变化事件；
- topic 删除、副本迁移、leader 切换等处理。

```
[2020-02-20 15:55:28,736][66] INFO [Controller id=1005] Partitions being reassigned: Map(reassign-test-0 -> Buffer(1001, 1003, 1005),
eassign-test-2 -> Buffer(1005, 1003, 1001), reassign-test-1 -> Buffer(1003, 1005, 1001)) (kafka.controller.KafkaController)
[2020-02-20 15:55:28,738][66] INFO [Controller id=1005] Currently active brokers in the cluster: Set(1005, 1001, 1006, 1007, 1003) (ka
ka.controller.KafkaController)
```

```
[2020-02-20 15:55:29,104][66] INFO [Controller id=1005] Partition reassign-test-0 to be reassigned is already assigned to replicas 1001
,1003,1005. Ignoring request for partition reassignment. (kafka.controller.KafkaController)
[2020-02-20 15:55:29,104][66] INFO [Controller id=1005] Handling reassignment of partition reassign-test-2 to new replicas 1005,1003,10
01 (kafka.controller.KafkaController)
[2020-02-20 15:55:29,106][66] INFO [Controller id=1005] New replicas 1005,1003,1001 for partition reassign-test-2 being reassigned not
yet caught up with the leader (kafka.controller.KafkaController)
[2020-02-20 15:55:29,110][66] INFO [Controller id=1005] Updated assigned replicas for partition reassign-test-2 being reassigned to 100
5,1003,1001,1006 (kafka.controller.KafkaController)
[2020-02-20 15:55:29,111][62] DEBUG [Controller id=1005] Updating leader epoch for partition reassign-test-2 (kafka.controller.KafkaCon
troller)
[2020-02-20 15:55:29,113][66] INFO [Controller id=1005] Updated leader epoch for partition reassign-test-2 to 1 (kafka.controller.Kafka
Controller)
[2020-02-20 15:55:29,116][66] INFO [Controller id=1005] Waiting for new replicas 1005,1003,1001 for partition reassign-test-2 being rea
ssigned to catch up with the leader (kafka.controller.KafkaController)
[2020-02-20 15:55:29,116][66] INFO [Controller id=1005] Handling reassignment of partition reassign-test-1 to new replicas 1003,1005,10
01 (kafka.controller.KafkaController)
[2020-02-20 15:55:29,117][66] INFO [Controller id=1005] New replicas 1003,1005,1001 for partition reassign-test-1 being reassigned not
yet caught up with the leader (kafka.controller.KafkaController)
[2020-02-20 15:55:29,118][66] INFO [Controller id=1005] Updated assigned replicas for partition reassign-test-1 being reassigned to 100
3,1005,1001,1006 (kafka.controller.KafkaController)
[2020-02-20 15:55:29,118][62] DEBUG [Controller id=1005] Updating leader epoch for partition reassign-test-1 (kafka.controller.KafkaCon
troller)
[2020-02-20 15:55:29,120][66] INFO [Controller id=1005] Updated leader epoch for partition reassign-test-1 to 1 (kafka.controller.Kafka
Controller)
[2020-02-20 15:55:29,121][66] INFO [Controller id=1005] Waiting for new replicas 1003,1005,1001 for partition reassign-test-1 being rea
ssigned to catch up with the leader (kafka.controller.KafkaController)
[2020-02-20 15:55:29,122][66] INFO [Controller id=1005] Removing partitions Set(reassign-test-0) from the list of reassigned partitions
 in zookeeper (kafka.controller.KafkaController)
```

重新验证进度，发现执行重分布完成了。

```
[root@ctl-nm-hhht-yxxya6-ceph-011 ~]# /usr/hdp/current/kafka-broker/bin/kafka-reassign-partitions.sh --zookeeper ctl-nm-hhht-yxxya6-cep
h-027.ctyuncdn.net:12181/kafka-auth-test-1 --verify --reassignment-json-file /data8/apps-dev/object-log/object-log-etlday/project_new.j
son
MYDIR=/usr/hdp/current/kafka-broker/bin/../bin
    broker.rack=CTG-DEFAULT
advertised.listeners=SASL_PLAINTEXT://ctl-nm-hhht-yxxya6-ceph-011.ctyuncdn.net:6667
Status of partition reassignment:
Reassignment of partition reassign-test-0 completed successfully
Reassignment of partition reassign-test-2 completed successfully
Reassignment of partition reassign-test-1 completed successfully
```