KIP-166 - Add a tool to make amounts of replicas and leaders on brokers balanced

Status

Current state: *Under Discussion*

Discussion thread: here

JIRA: ⚠ Unable to render Jira issues macro, execution error.

Please keep the discussion on the mailing list rather than commenting on the wiki (wiki discussions get unwieldy fast).

Motivation

I think kafka need to add a tool to balance amounts of replicas and leaders on every brokers .Current kafka server does not provide a function to balance cluster .When we use console command to create a topic without 'replicaAssignmentOpt' ,kafka server assigns replicas by a "replicaAssignment" function . The function spread the replicas evenly among brokers to make cluster balanced.It can not make sure the cluster is balanced.Even though it is balanced, the balanced state will be destroyed.e.g.

1.Adding some new broker to current kafka cluster.There is nothing in a new broker.The amounts of replicas and leaders in new brokers are always fewer than old ones.

2.Using ReassignPartitionsCommand will change the replica and leader balance of cluster.e.g.

A reassignment-json-file as :

```
{ "partitions":
     [{"topic": "topic-test",
       "partition": 0,
       "replicas": [1,2,3] },
     {"topic": "topic-test",
       "partition": 1,
       "replicas": [1,2,3] },
      {"topic": "topic-test",
       "partition": 2,
       "replicas": [1,2,3] }
    ],
     "version":1
    }
```

It means some one want to assign replicas of 3 partitions to broker 1,2,3.The one was accustomed to write AR in ascending order.Kafka server choose a PreferredReplicaLeader which be put at the first position of the AR to balance leader .If all leaders are the PreferredReplicaLeader of AR ,then kafka server thinks leaders in cluster is balanced.In above case,kafka server thinks it is balance when all leaders are replica 1.In fact the leader balance is not good.

So it is necessary to add to add a tool to deal with replica and leader balance.

Public Interfaces

Add a new tool.There are 6 options in the new tool.

--leader-balance	to execute leader balance
--replica-balance	to execute replica balance
--cluster-balance	to execute replica balance and leader balance
--show-cluster-sate	to show the state of balance
--zookeeper	zookeeper urls <horst:port>
--replica-balance-threshold-percentage	a threshold parameter

Proposed Changes

We use an algorithm to get a balanced assignment of replicas or leaders,and call ReassignPartitionsCommand to let kafka sever move replicas or change the position of replicas in ARs.

We designed 4 command of this tools:

- show-cluster-state   This command prints out the state of cluster .(amount of replicas and leaders on every brokers)
- leader-balance   This command makes amounts of leaders on every brokers balanced
- replica-balance   This command makes amounts of replicas on every brokers balanced
- cluster-balance   This command makes amounts of replicas and leaders on every brokers balanced

1.show-cluster-state

Get information of every TP (topic and partition) from zookeeper,and calculate the amounts of replicas and "PreferredReplicaLeader" on every alive brokers.Print out them to help user

to assess whether do a balance.

2.leader-balance

Get information of every TP (topic and partition) from zookeeper,and then get the state of "PreferredReplicaLeader" on every alive brokers.Calculate the avg of amount of"PreferredReplicaLeader" on every brokers.We think a leader balance state is the amount of preferred leaders of each brokers are close to the avg. We call brokers which need to move leader position "SourceBroker",and brokers need to get we call them "TargetBroker".During moving ,we should make sure the amount of leaders on SourceBroker is greater than the avg,the amount of leaders on TargetBroker is less than the avg, and the TargetBroker should has the same TP of the leader SourceBroker want to move.

3.replica-balance

Replica balance is similar to leader balance.The different is during moving the replica we should make sure there is not a replica which is belong to the same TP with the replica SourceBroker want to move on TargetBroker.   We also provide a "threshold" option .Within the "threshold" option we think it is balanced.Because replica balance costs more than leader balance.It touches on data migration.The default is 10 which means 10% of the avg.

4.cluster-balance

Cluster-balance do replica balance first,and then execute leader balance ,finally use ReassignPartitionsCommand to assign partition with the balance assignment we get above.

Click here to view the code.

Compatibility, Deprecation, and Migration Plan

- This is a new tool. It will not cause any compatibility issues.

Rejected Alternatives

We can also use ReassignPartitionsCommand to move replicas to do replica balance and change the first replica in ARs to do leader balance by manual operation.However it will become more and more difficult with topic and broker increasing.