KIP-570: Add leader epoch in StopReplicaRequest

Status

Current state: *Accepted [One of "Under Discussion", "Accepted", "Rejected"]*

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

Unlike the LeaderAndIsrRequest, the StopReplicaRequest does not include the leader epoch which makes it vulnerable to reordering. This KIP proposes to add the leader epoch for each partition in the StopReplicaRequest and the broker will verify the epoch before proceeding with the StopReplicaRequest.

Public Interfaces

We will bump the version of the StopReplicaRequest/StopReplicaResponse and add the leader epoch for each partition in the request.

```
{
  "apiKey": 5,
  "type": "request",
  "name": "StopReplicaRequest",
  // Version 1 adds the broker epoch and reorganizes the partitions to be stored
  // per topic.
  //
  // Version 2 is the first flexible version.
  //
  // Version 3 reorganizes the partitions to be stored, adds the leader epoch
  // and the delete partition fields per partition (KIP-570).
  "validVersions": "0-3",
  "flexibleVersions": "2+",
  "fields": [
    { "name": "ControllerId", "type": "int32", "versions": "0+", "entityType": "brokerId",
      "about": "The controller id." },
    { "name": "ControllerEpoch", "type": "int32", "versions": "0+",
      "about": "The controller epoch." },
    { "name": "BrokerEpoch", "type": "int64", "versions": "1+", "default": "-1", "ignorable": true,
      "about": "The broker epoch." },
    { "name": "DeletePartitions", "type": "bool", "versions": "0-2",
      "about": "Whether these partitions should be deleted." },
    { "name": "UngroupedPartitions", "type": "[]StopReplicaPartitionV0", "versions": "0",
      "about": "The partitions to stop.", "fields": [
      { "name": "TopicName", "type": "string", "versions": "0", "entityType": "topicName",
        "about": "The topic name." },
      { "name": "PartitionIndex", "type": "int32", "versions": "0",
        "about": "The partition index." }
    ]},
    { "name": "Topics", "type": "[]StopReplicaTopicV1", "versions": "1-2",
      "about": "The topics to stop.", "fields": [
      { "name": "Name", "type": "string", "versions": "1-2", "entityType": "topicName",
        "about": "The topic name." },
      { "name": "PartitionIndexes", "type": "[]int32", "versions": "1-2",
        "about": "The partition indexes." }
    ]},
        // New Structure from V3 on
    { "name": "TopicStates", "type": "[]StopReplicaTopicState", "versions": "3+",
      "about": "Each topic.", "fields": [
      { "name": "TopicName", "type": "string", "versions": "3+", "entityType": "topicName",
        "about": "The topic name." },
      { "name": "PartitionStates", "type": "[]StopReplicaPartitionState", "versions": "3+",
        "about": "The state of each partition", "fields": [
        { "name": "PartitionIndex", "type": "int32", "versions": "3+",
          "about": "The partition index." },
        { "name": "LeaderEpoch", "type": "int32", "versions": "3+", "default": "-1",
          "about": "The leader epoch." },
        { "name": "DeletePartition", "type": "bool", "versions": "3+",
          "about": "Whether this partition should be deleted." }
      ]}
    ]}
  ]
}
```

Proposed Changes

The controller will include the leader epoch of each partition when sending out an StopReplicaRequest. The broker will verify the epoch of each partitions and send an `FENCED_LEADER_EPOCH` error when the leader epoch received is older than the known one. When a topic is deleted, the leader epoch is not bumped. In this case, we will send a sentinel (-2) which overrides any existing epoch. Older version of the request will use a sentinel (-1) to indicate the the leader epoch is not present when the controller is still on the old version during the upgrade.

Starting from V3, only one StopReplica request will be sent by the controller, combining the partitions to be deleted and the partitions to stopped only.

Compatibility, Deprecation, and Migration Plan

The change is backward compatible with older broker.

Rejected Alternatives

N/A