KIP-461 - Improve Replica Fetcher behavior at handling partition failure

Status

Current state: *Accepted*

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

The replica fetcher threads handle multiple partitions. In case a partition fails, the replica fetcher thread associated with that partition terminates. The partitions that have caught up and are running well are also left untracked with termination of the thread which leads to under-replicated partitions. A better approach would be, whenever a partition crashes, the concerned thread should stop tracking the crashed one and continue handling rest of the partitions.

Public Interfaces

New metrics:

- FailedPartitionsCount - Count of partitions that have failed. Instead of separate metrics, clientId is used as a tag to distinguish between *Replica* and *ReplicaAlterLogDir* fetchers, keeping it consistent with metric like MaxLag.

Proposed Changes

In case a partition fails, the replica fetcher thread would stop tracking the failed partition. A set failedPartitions would be used to keep a track of it. Instead of throwing an exception which ends up terminating the thread, an error message will be logged and the partition will be added to the failedPartitions set. The partition would be removed from the fetcherLagStats and partitionStates since partition lag cannot be accurately tracked once fetching is stopped. The thread would continue monitoring rest of the partitions which are lost in the current scenario.

If all partitions for a fetcher thread are marked as failed, the thread would be shut down. In cases where a replica is deleted on a broker through a StopReplicaRequest while the partition is present in failedPartitions set, the partition would be removed from the set.

Until the next leader epoch, the partition would remain in the failedPartitions set. At the leader epoch, the failed partitions would be marked as un-failed by removing from the set for failed partitions. Hereafter, the controller can choose the partition as leader or follower and would follow the usual behavior.

This logic will be implemented in AbstractFetcherThread so that it applies to both replica and log dir fetchers.

Compatibility, Deprecation, and Migration Plan

- The metric FailedPartitionCount would keep track of the failed partitions. It's a newly added metric which would help keep track of failed partitions. It would avoid losing several healthy partitions in case partition failure occurs.

Rejected Alternatives

- Retries - The thread can make attempts to connect to the failed partition which would mostly hit the same problem.
- Shutting down the broker - If more than 50% partitions on a broker have failed, the broker can be shut down, left as a potential future work.