Item 118
git_comments:

git_commits:

1. **summary:** [SPARK-23438][DSTREAMS] Fix DStreams data loss with WAL when driver crashes
   message: [SPARK-23438][DSTREAMS] Fix DStreams data loss with WAL when driver crashes There is a race condition introduced in SPARK-11141 which could cause data loss. The problem is that ReceivedBlockTracker.insertAllocatedBatch function assumes that all blocks from streamIdToUnallocatedBlockQueues allocated to the batch and clears the queue. In this PR only the allocated blocks will be removed from the queue which will prevent data loss. Additional unit test + manually. Author: Gabor Somogyi <gabor.g.somogyi@gmail.com> Closes #20620 from gaborgsomogyi/SPARK-23438. (cherry picked from commit b308182f233b8840dfe0e6b5736d2f2746f40757) Signed-off-by: Marcelo Vanzin <vanzin@cloudera.com>

github_issues:

github_issues_comments:

github_pulls:

1. **title:** [SPARK-23438][DSTREAMS] Fix DStreams data loss with WAL when driver crashes
   body: ## What changes were proposed in this pull request? There is a race condition introduced in SPARK-11141 which could cause data loss. The problem is that ReceivedBlockTracker.insertAllocatedBatch function assumes that all blocks from streamIdToUnallocatedBlockQueues allocated to the batch and clears the queue. In this PR only the allocated blocks will be removed from the queue which will prevent data loss. ## How was this patch tested? Additional unit test + manually.

github_pulls_comments:

1. @jose-torres @tdas @zsxwing could you take a look at this please?
2. ok to test
3. **[Test build #87489 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/87489/testReport)** for PR 20620 at commit [`152fec4`] (https://github.com/apache/spark/commit/152fec431218161e538c377a6cb82753100dc70b). * This patch **fails Spark unit tests**. * This patch merges cleanly. * This patch adds no public classes.
4. **[Test build #87494 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/87494/testReport)** for PR 20620 at commit [`bd46d1c`] (https://github.com/apache/spark/commit/bd46d1cb63e7a04e0236f7b1bf70b46fb55f3ea4). * This patch passes all tests. * This patch merges cleanly. * This patch adds no public classes.
5. **[Test build #87519 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/87519/testReport)** for PR 20620 at commit [`23e0204`] (https://github.com/apache/spark/commit/23e020438c851502522f2328f01728e43c1fba99). * This patch **fails due to an unknown error code, -9**. * This patch merges cleanly. * This patch adds no public classes.
6. Seems like unrelated issue.
7. retest this please.
8. **[Test build #87522 has finished] (https://amplab.cs.berkeley.edu/jenkins/job/SparkPullRequestBuilder/87522/testReport)** for PR 20620 at commit [`23e0204`] (https://github.com/apache/spark/commit/23e020438c851502522f2328f01728e43c1fba99). * This patch passes all tests. * This patch merges cleanly. * This patch adds no public classes.
9. LGTM.
10. Merging to master, will try back to 2.0.

11. (Argh, the wifi here is horrible, I'll need to manually merge things, so hang on a sec...)

github_pulls_reviews:

1. nit: Can we use another name (maybe `allocatedBlocksInStream`?) other than `allocatedBlocks` to avoid confusion?
2. Sure, fixed.

jira_issues:

jira_issues_comments: