Item 43
git_comments:

git_commits:

1. **summary:** This closes #2456
   message: This closes #2456

github_issues:

github_issues_comments:

github_pulls:

1. **title:** [BEAM-778] Fix the Compressed file seek tests on windows
   body: Be sure to do all of the following to help us incorporate your contribution quickly and easily: - [ ] Make sure the PR title is formatted like: `[BEAM-<Jira issue #>] Description of pull request` - [ ] Make sure tests pass via `mvn clean verify`. (Even better, enable Travis-CI on your fork and ensure the whole test matrix passes). - [ ] Replace `<Jira issue #>` in the title with the actual Jira issue number, if there is one. - [ ] If this contribution is large, please file an Apache [Individual Contributor License Agreement] (https://www.apache.org/licenses/icla.pdf). --- R: @chamikaramj @tibkiss The seek tests were failing on windows due to two main reasons: (1) files not being closed before removing and (2) files not being opened as binary files.

github_pulls_comments:

1. Refer to this link for build results (access rights to CI server needed): https://builds.apache.org/job/beam_PreCommit_Java_MavenInstall/9263/<h2>Build result: FAILURE</h2>[...truncated 2.55 MB...] at java.lang.Thread.run(Thread.java:745)Caused by: org.apache.maven.plugin.MojoExecutionException: Command execution failed. at org.codehaus.mojo.exec.ExecMojo.execute(ExecMojo.java:302) at org.apache.maven.plugin.DefaultBuildPluginManager.executeMojo(DefaultBuildPluginManager.java:134) at org.apache.maven.lifecycle.internal.MojoExecutor.execute(MojoExecutor.java:208) ... 31 moreCaused by: org.apache.commons.exec.ExecuteException: Process exited with an error: 1 (Exit value: 1) at org.apache.commons.exec.DefaultExecutor.executeInternal(DefaultExecutor.java:404) at org.apache.commons.exec.DefaultExecutor.execute(DefaultExecutor.java:166) at org.codehaus.mojo.exec.ExecMojo.executeCommandLine(ExecMojo.java:764) at org.codehaus.mojo.exec.ExecMojo.executeCommandLine(ExecMojo.java:711) at org.codehaus.mojo.exec.ExecMojo.execute(ExecMojo.java:289) ... 33 more2017-04-07T00:22:09.485 [ERROR] 2017-04-07T00:22:09.485 [ERROR] Re-run Maven using the -X switch to enable full debug logging.2017-04-07T00:22:09.485 [ERROR] 2017-04-07T00:22:09.485 [ERROR] For more information about the errors and possible solutions, please read the following articles:2017-04-07T00:22:09.485 [ERROR] [Help 1] http://cwiki.apache.org/confluence/display/MAVEN/MojoExecutionException2017-04-07T00:22:09.486 [ERROR] 2017-04-07T00:22:09.486 [ERROR] After correcting the problems, you can resume the build with the command2017-04-07T00:22:09.486 [ERROR] mvn <goals> -rf :beam-sdks-pythonchannel stoppedSetting status of ee225fa2c9489daa8c4b650f2a30d5d55b0a1f3d to FAILURE with url https://builds.apache.org/job/beam_PreCommit_Java_MavenInstall/9263/ and message: 'Build finished. 'Using context: Jenkins: Maven clean install --none--
2. [![Coverage Status](https://coveralls.io/builds/10973850/badge)](https://coveralls.io/builds/10973850) Changes Unknown when pulling **a4c795fbd78a6531971f586ca1f0c76ecf49aa9d on sb2nov:BEAM-windows-test-fix** into ** on apache:master**.
3. Refer to this link for build results (access rights to CI server needed): https://builds.apache.org/job/beam_PreCommit_Java_MavenInstall/9264/ --none--
4. Nice catch @sb2nov. It was clearly a mistake from my side not testing BEAM-778 on Windows. Your fix looks reasonable to me from the content perspective (I don't have windows machine around to test it). Two thoughts: * Is it possible to setup Windows Jenkins Slave/Travis for testing? * Currently the tests are using `with open('filename', mode)`. This could be further simplified to `with CompressedFile(open(..),..)`. That way we would also test the `__enter__` and `__exit__` methods of CompressedFile.

5. LGTM. Thanks for fixing this.
6. @chamikaramj Thanks for merging. @tibkiss Makes sense. (1) I think Jason is looking into adding a windows slave. We could also try to investigate using something like https://docs.tea-ci.org/usage/overview/ as it is free for open source projects (2) I'll followup with that in a PR shortly.
7. cc: @jasonkuster for the Windows testing question.

github_pulls_reviews:

jira_issues:

1. **summary:** Make filesystem._CompressedFile seekable.
   description: We have a TODO to make filesystem._CompressedFile seekable. https://github.com/apache/incubator-beam/blob/python-sdk/sdks/python/apache_beam/io/fileio.py#L692 Without this, compressed file objects produce for FileBasedSource implementations may not be able to use libraries that utilize methods seek() and tell(). For example tarfile.open().

jira_issues_comments:

1. [~chamikara]: I'd like to work on this improvement. Could you please assign it to me? Thanks!
2. Thanks for looking into this issue. I couldn't find you in the list of JIRA users. [~davor] could you assign this issue to Tibor.
3. Welcome [~tibor.kiss@gmail.com]! Thanks for your contribution!
4. [~davor]: Thanks for the privs & assignment!
5. The implementation today maintains a local {{_read_buffer}} object which is used all the way on the read path. I suspect that the _read_buffer is created to bridge the gap between zlib module's functionality (provides only block decompress and compress) and the required operations of the file object (read bytes and read line). My impression is that if we would replace zlib module with gzip (which builds on top of gzip) then we could simply bridge the read operations to gzip's respective methods without the need of having local buffer. Bzip2 module also supports read operations. Bonus would be that seek() functionality would come for 'free' as both gzip and bzip2 supports seek() and tell(). [~robertwb] / [~sbilac] / [~katsiapis@google.com] / [~altay]: Wondering if you considered using gzip module? What are your thoughts on ditching read_buffer by bridged file ops to bzip2/gzip?
6. Yes, when I had looked at that in the past, it was very easy to use gzip.GzipFile(..., fileobj = self._file, ...) [1] but unfortunately not so for Bzip2 [2] or Snappy [3]. And we wanted to share as much implementation as possible (as opposed to have completely different codepaths for each compression type). Provided that we can have a single interface that allows us to handle Gzip/Bzip2 (and ideally in the future Snappy and other whole-file compression techniques) with minimal diffs, changing the underlying implementation is I think fair game. [1] https://docs.python.org/2/library/gzip.html#gzip.GzipFile and https://github.com/apache/beam/blob/master/sdks/python/apache_beam/io/filesystem.py#L103 [2] https://docs.python.org/2/library/bz2.html#bz2.BZ2File [3] https://pypi.python.org/pypi/python-snappy
7. **body:** [~katsiapis@google.com]: Thanks for the explanation! BZ2 has certainly mislead me: fileobj support is only available in Python 3.x. It is not possible to implement fileobj support by extending the class under Python 2.7 as BZ2 is written as a C module and uses FILE* inside. Therefore I'll stick with the original design and implement seek() on top of that.
   label: code-design
8. I'm still working on seek() implementation and I have noticed that there is no lock to protect the {{_read_buffer}} object. I'm not completely sure if it is a valid scenario that multiple threads accessing the same _CompressedFile object though. Any thoughts on extending this class with a lock on {{_read_buffer}}?
9. **body:** In general the Beam programming model usually requires thread-compatibility (not the stronger thread-safety). I would hazard a guess that we should follow suite for _CompressedFile (ie no need for locking?). Having said that it might be worth documenting the thread-compatibility but [~chamikara] or [~altay] would have a better sense about overall documentation etc.
    label: documentation
10. **body:** Currently this is not an issue since Beam FileBasedSoure and FileBasedSink are the only users of CompressedFile/File objects and they are used in a pretty straightforward way where each FileBasedSource/FileBasedSink object owns it's File/CompressedFile object and reading is done using a single thread. A secondary thread that performs dynamic work rebalancing might execute seek() operations for File objects but not for CompressedFile objects. In the future we might have other places

where we access CompressedFile objects using multiple thread but I think we should probably wait till such needs arise. Also it might be enough to declare CompressedFile objects to be not thread safe and expect users to address thread safety instead of embedding a lock in CompressedFile objects which would potentially add a performance penalty for all users. WDYT ?
label: requirement

11. **body:** Thanks for the insights, [~katsiapis@google.com] & [~chamikara]. As there is no immediate risk of concurrency issue I'd also opt for not extending _CompressedFile with a lock prematurely. I'll make a comment about the (lack of) thread safety in _CompressedFile in the PR associated with this JIRA.
label: requirement
12. Changed the title from 'Make fileio._CompressedFile seekable.' to 'Make filesystem._CompressedFile seekable.' as BEAM-1441 moved _CompressedFile from fileio module to filesystem.
13. GitHub user tibkiss opened a pull request: https://github.com/apache/beam/pull/2392 [BEAM-778] Make filesystem._CompressedFile seekable. Be sure to do all of the following to help us incorporate your contribution quickly and easily: - [ ] Make sure the PR title is formatted like: `[BEAM-<Jira issue #>] Description of pull request` - [ ] Make sure tests pass via `mvn clean verify`. (Even better, enable Travis-CI on your fork and ensure the whole test matrix passes). - [ ] Replace `<Jira issue #>` in the title with the actual Jira issue number, if there is one. - [ ] If this contribution is large, please file an Apache [Individual Contributor License Agreement](https://www.apache.org/licenses/icla.txt). --- You can merge this pull request into a Git repository by running: $ git pull https://github.com/tibkiss/beam BEAM-778 Alternatively you can review and apply these changes as the patch at: https://github.com/apache/beam/pull/2392.patch To close this pull request, make a commit to your master/trunk branch with (at least) the following in the commit message: This closes #2392 ---- commit a11859d164775eec6e2c870138a7d883db47faf2 Author: Tibor Kiss <tibor.kiss@gmail.com> Date: 2017-03-31T05:11:07Z [BEAM-778] Make filesystem._CompressedFile seekable. ----
14. Notes: - Seeks are supported in files which are opened for read. Seek in write/append file modes are not supported. - BEAM-1441 has moved _CompressedFile to filesystem.py, but left the testcases in fileio_test.py. This has been corrected in this PR too.
15. Github user asfgit closed the pull request at: https://github.com/apache/beam/pull/2392
16. GitHub user sb2nov opened a pull request: https://github.com/apache/beam/pull/2456 [BEAM-778] Fix the Compressed file seek tests on windows Be sure to do all of the following to help us incorporate your contribution quickly and easily: - [ ] Make sure the PR title is formatted like: `[BEAM-<Jira issue #>] Description of pull request` - [ ] Make sure tests pass via `mvn clean verify`. (Even better, enable Travis-CI on your fork and ensure the whole test matrix passes). - [ ] Replace `<Jira issue #>` in the title with the actual Jira issue number, if there is one. - [ ] If this contribution is large, please file an Apache [Individual Contributor License Agreement](https://www.apache.org/licenses/icla.pdf). --- R: @chamikaramj @tibkiss The seek tests were failing on windows due to two main reasons: (1) files not being closed before removing and (2) files not being opened as binary files. You can merge this pull request into a Git repository by running: $ git pull https://github.com/sb2nov/beam BEAM-windows-test-fix Alternatively you can review and apply these changes as the patch at: https://github.com/apache/beam/pull/2456.patch To close this pull request, make a commit to your master/trunk branch with (at least) the following in the commit message: This closes #2456 ---- commit ee225fa2c9489daa8c4b650f2a30d5d55b0a1f3d Author: Sourabh Bajaj <sourabhbajaj@google.com> Date: 2017-04-06T23:23:33Z [BEAM-778] Fix the Compressed file seek tests on windows ----
17. Github user asfgit closed the pull request at: https://github.com/apache/beam/pull/2456
18. This can be closed now.
19. For some reason I cannot close/resolve this. [~chamikara]: Could you please resolve this issue? Thanks!
20. Done. Thanks.