Item 45
git_comments:

git_commits:

1. **summary:** [BEAM-11] add Spark runner to included runners
   message: [BEAM-11] add Spark runner to included runners

github_issues:

github_issues_comments:

github_pulls:

github_pulls_comments:

github_pulls_reviews:

jira_issues:

1. **summary:** Integrate Spark runner with Beam
   description: Refactor and integrate the Spark runner code against Google's contributed version of Dataflow - Beam.

jira_issues_comments:

1. I have a compiling-but-untested runner on my local against dataflow-1.5.0-SNAPSHOT (which will become the first drop of Beam any day now) and flink-1.0-SNAPSHOT. I intend to at least check it still runs.
2. Great! If the first SNAPSHOT of Beam doesn't break the runner then once the runner code is imported we'll close this ticket.
3. It was alas very broken but I think it's back together again. I'm in the dataflow team at Google so I'll track till the first drop then publish the cl.
4. BTW I'd crossed wires in the above - I'm working on the flink runner not the spark runner.
5. **body:** Not sure if others are working on this, but the commit linked below is probably the smallest possible change to get spark-dataflow running with the current Beam code. Here it is: https://github.com/rbrush/spark-dataflow/commit/0a11d747eeb6bb47bb46e179deca4c85a9d5cf33 We need to do quite a bit more with the runner before it's broadly usable; see the ugly "TODO" around state internals in the commit. So perhaps the best path forward is to just create a development branch of Beam that includes the dataflow runner and we can improve on it there? Once it's in a better state we can squash/rebase (or whatever conventions this project follows) to get a clean merge into master. I'm happy to create the branch if desired (although I lack commit privs), or feel free to just grab the code from the above commit if it makes sense.
   label: code-design
6. Yes, I'm on it :) pending the code drop. Hang tight.. I'll take a look at your work, and compare with mine. I'd merge our work, or tell you to do a pull request, but this is a weird time since there is still no code... But no worries, the code will be dropped soon and we can get things going. There is a lot of work to do on supporting the Beam model, especially in streaming, but also in metric reporting and more, as stated in the technical vision document: https://drive.google.com/folderview?id=0B-IhJZh9Ab52OFBVZHpsNjc4eXc&usp=sharing Currently, the runner supports batch processing, and some, limited, stream processing, which is OK - we can state the runner's current capabilities (keep in mind that not all runners will support the entire model). Having said that, we will add more support as we go. Concerning branch organization see: https://drive.google.com/folderview?id=0B-IhJZh9Ab52OFBVZHpsNjc4eXc&usp=sharing It makes total sense to develop large features of the runner in a feature branch, and merge once done.
7. Sounds great, Amit. No worries either way on the code; I had just been playing with getting the Spark runner going over the current code at https://github.com/apache/incubator-beam and shared just in case it might be useful. I'm confident that we can move the runner forward once the code merged. (I assume we should track BEAM-6 for that?)

8. It sounds good. I will try to ping Tom today. For the directories structure, [~davor] said that we can push as it is and re-organize the directories in a second step (it's what I'm doing: just renaming and legal now in a first step).
9. Github user asfgit closed the pull request at: https://github.com/apache/incubator-beam/pull/42
10. Closed by PR-42