# Become a contributor

Making Changes, Running Patched Pipeline, and Contributing back to Beam

Yi Hu
Google
GitHub: @Abacn

BEAM SUMMIT
NYC 2025

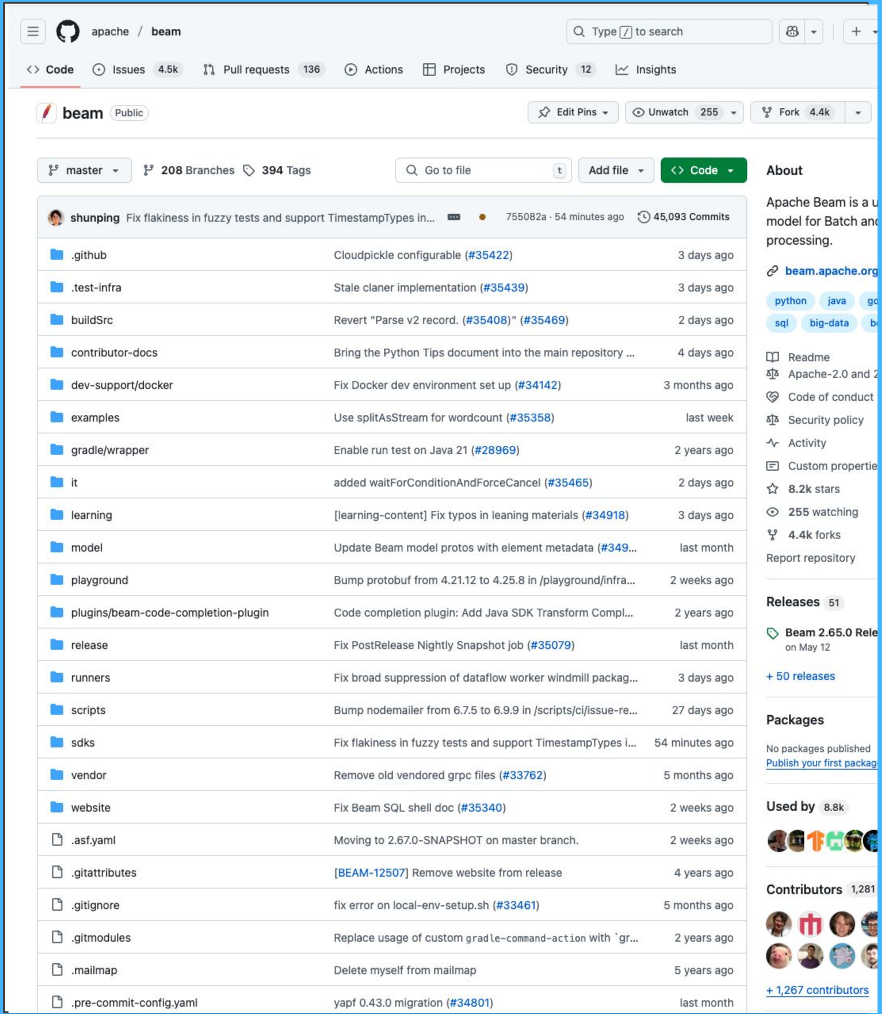
# Agenda

- Beam repo
- Build and run patched Beam
  - Development Setup
  - Run patched pipeline
- Contributing Back!
  - Pull Requests
- When do I expect my change get released
- Recap

# Apache Beam Repo

*Pretty much* a “monorepo” for Beam.

- **/sdks/java** Java SDK
  - **/sdks/java/core** Java core
  - **/sdks/java/harness** SDK harness (entrypoint of SDK container)
  - ...
- **/runners** Java runner supports
  - **/runners/google-cloud-dataflow-java** Dataflow runner (job submission, translation, etc)
  - **/runners/google-cloud-dataflow-java/worker** Worker on **Dataflow legacy runner**



shunping Fix flakiness in fuzzy tests and support TimestampTypes in... 755082a · 54 minutes ago 45,093 Commits

| | | |
|---|---|---|
| .github | Cloudpickle configurable (#35422) | 3 days ago |
| .test-infra | Stale claner implementation (#35439) | 3 days ago |
| buildSrc | Revert "Parse v2 record. (#35408)" (#35469) | 2 days ago |
| contributor-docs | Bring the Python Tips document into the main repository ... | 4 days ago |
| dev-support/docker | Fix Docker dev environment set up (#34142) | 3 months ago |
| examples | Use splitAsStream for wordcount (#35358) | last week |
| gradle/wrapper | Enable run test on Java 21 (#28969) | 2 years ago |
| it | added waitForConditionAndForceCancel (#35465) | 2 days ago |
| learning | [learning-content] Fix typos in leaning materials (#34918) | 3 days ago |
| model | Update Beam model protos with element metadata (#349... | last month |
| playground | Bump protobuf from 4.21.12 to 4.25.8 in /playground/infra... | 2 weeks ago |
| plugins/beam-code-completion-plugin | Code completion plugin: Add Java SDK Transform Compl... | 2 years ago |
| release | Fix PostRelease Nightly Snapshot job (#35079) | last month |
| runners | Fix broad suppression of dataflow worker windmill packag... | 3 days ago |
| scripts | Bump nodemailer from 6.7.5 to 6.9.9 in /scripts/ci/issue-re... | 27 days ago |
| sdks | Fix flakiness in fuzzy tests and support TimestampTypes i... | 54 minutes ago |
| vendor | Remove old vendored grpc files (#33762) | 5 months ago |
| website | Fix Beam SQL shell doc (#35340) | 2 weeks ago |
| .asf.yaml | Moving to 2.67.0-SNAPSHOT on master branch. | 2 weeks ago |
| .gitattributes | [BEAM-12507] Remove website from release | 4 years ago |
| .gitignore | fix error on local-env-setup.sh (#33461) | 5 months ago |
| .gitmodules | Replace usage of custom `gradle-command-action` with `gr... | 2 years ago |
| .mailmap | Delete myself from mailmap | 5 years ago |
| .pre-commit-config.yaml | yapf 0.43.0 migration (#34801) | last month |

About

Apache Beam is a u
model for Batch and
processing.

beam.apache.org

python java go
sql big-data b

Readme
Apache-2.0 and 2
Code of conduct
Security policy
Activity
Custom propertie
8.2k stars
255 watching
4.4k forks
Report repository

Releases 51

Beam 2.65.0 Rele
on May 12

+ 50 releases

Packages

No packages published
Publish your first packag

Used by 8.8k

Contributors 1,281

+ 1,267 contributors

# Apache Beam Repo

*Pretty much* a “monorepo” for Beam.

- **/sdks/java** Java SDK
  - **/sdks/java/core** Java core
  - **/sdks/java/harness** SDK harness (entrypoint of SDK container)
  - ...
- **/runners** Java runner supports
  - **/runners/google-cloud-dataflow-java** Dataflow runner (job submission, translation, etc)
  - **/runners/google-cloud-dataflow-java/worker** Worker on **Dataflow legacy runner**

- **/sdks/python** Python SDK
- **/sdks/go** Go SDK
- **/model** Beam proto definitions (proto SDK)
- **/buildSrc** Gradle Build plugin, including
  - Global Java dependency managements
  - Common test logics
  - Java version handling, etc
- **/website** Beam Websites
- **/.github/workflows** GitHub workflows

Starter repo:

https://github.com/apache/beam-starter-java
https://github.com/apache/beam-starter-python
https://github.com/apache/beam-starter-go

# Apache Beam repo: Build Infra

- A huge Gradle plugin *buildSrc/src/main/groovy/org/apache/beam/gradle/BeamModulePlugin* manages everything. For example, creating a new java (sub)project with

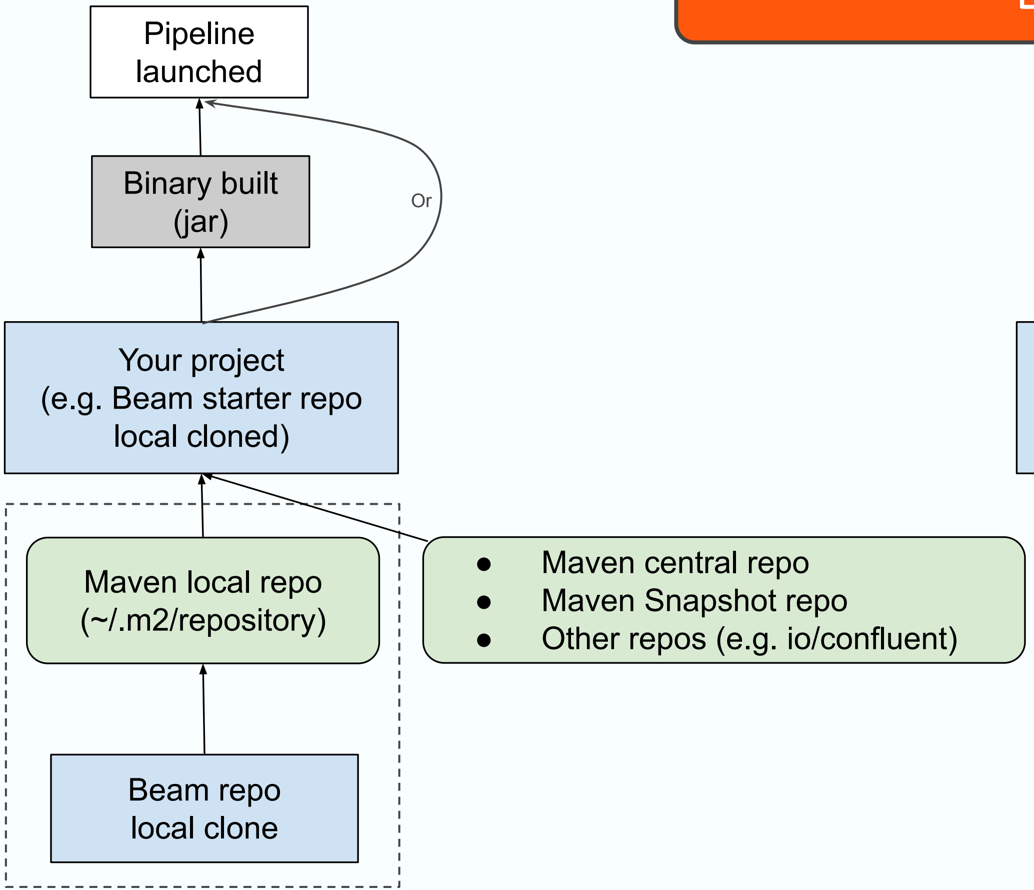
```
apply plugin: 'org.apache.beam.module'

applyJavaNature( ... )
```

Relevant usage of BeamModulePlugin:

- Manage Java dependencies
- Configuring projects (Java, Python, Go, Proto, Docker, Avro, etc)
    - java -> applyJavaNature
    - Define common custom tasks for each type of projects
        - test : run unit tests
        - spotlessApply : format tests

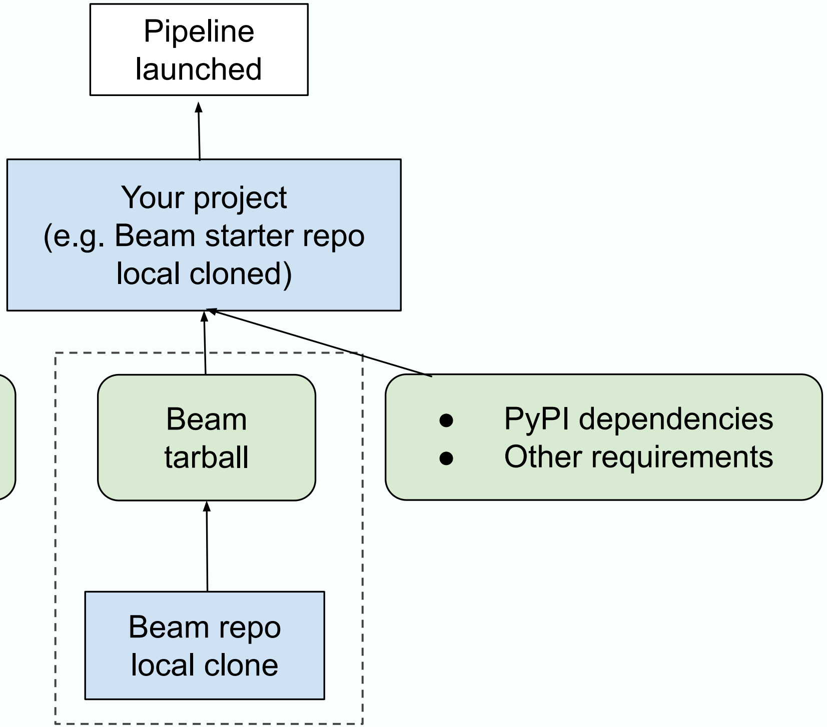
# Build and run patched Beam

- Java Pipeline



Pipeline launched
Binary built (jar)
Or
Your project (e.g. Beam starter repo local cloned)
Maven local repo (~/.m2/repository)
Maven central repo
Maven Snapshot repo
Other repos (e.g. io/confluent)
Beam repo local clone

- Python Pipeline



# Environment setup

Code change guide: https://github.com/apache/beam/blob/master/contributor-docs/code-change-guide.md

Prerequisites. In your PATH, having (as of July 2025)

- A Java environment (java11 preferably, java17 or 21 should also work)
  Recommended: Use **sdkman** to manage Java versions (https://sdkman.io/install)
- A Python environment (py39-py312 are supported)
  Recommended: Use **pyenv** and **virtualenv** to manage Python versions
- A Go environment

Optional:

- Docker environment is now optional. Alternatives:
  - Portable options --environment_type==LOOPBACK
  - Dataflow runner v2 job: --experiment=use_dataflow_worker_jar (#33508)

# IDE setup

Java: IntelliJ Setup

- https://www.jetbrains.com/idea/download/ community version should suffice for OSS
- From IntelliJ, open /beam (Beam root, instead of sdks/java)
- Wait for indexing (minutes)

It should just work (after recent fix #35167), as gradle is a self contained build tool

- find examples/java/build.gradle, click "run" button at task wordCount
  [Note] Running examples directly on Beam repo now support different runners (#35262)

Python/Go: can use VSCode

- Open sdks/python or sdks/go folder
- Setup interpreter

Further reading: https://github.com/apache/beam/blob/master/contributor-docs/code-change-guide.md

# Run patched Java pipeline

Pipeline implemented in a user project (outside beam repo)

1. Modify Beam version
   a. Maven project: pom.xml
   b. Gradle project: build.gradle
2. (if patched Beam Head) Add snapshot repository
   a. Maven project pom.xml

```
<repository>
      <id>Maven-Snapshot</id>
      <name>maven snapshot repository</name>
      <url>https://repository.apache.org/content/groups/snapshots/</url>
    </repository>
```

   b. Gradle project build.gradle

```
repositories {
    mavenLocal()
    mavenCentral()
    maven {
        url 'https://repository.apache.org/content/groups/snapshots'
    }
}
```

# Run patched Java pipeline

In Beam side

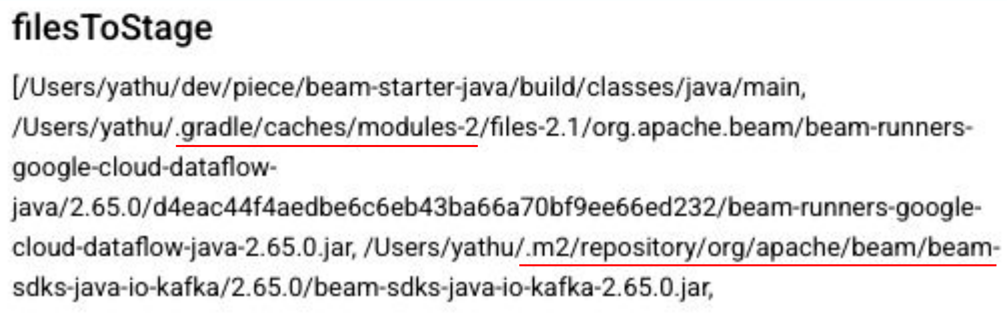
1. Compile the project involving the code change with (e.g. if modified sdks/java/io/kafka)

```
./gradlew -Ppublishing -p sdks/java/io/kafka publishToMavenLocal
```

This will publish the artifact with modified code to Maven Local repository (~/.m2/repository) by default, and **it will be picked up when executing user pipeline**

After patched Beam artifact built, add Beam dependency, build and submit your pipeline the same way as before.

To check the effectiveness on Dataflow UI:



**filesToStage**

[/Users/yathu/dev/piece/beam-starter-java/build/classes/java/main, /Users/yathu/.gradle/caches/modules-2/files-2.1/org.apache.beam/beam-runners-google-cloud-dataflow-java/2.65.0/d4eac44f4aedbe6c6eb43ba66a70bf9ee66ed232/beam-runners-google-cloud-dataflow-java-2.65.0.jar, /Users/yathu/.m2/repository/org/apache/beam/beam-sdks-java-io-kafka/2.65.0/beam-sdks-java-io-kafka-2.65.0.jar,

# Run patched Python Pipeline

1. Build the Beam SDK tarball. Under sdks/python, run python -m build --sdist.
2. Under current virtualenv, pip install /path/to/apache-beam.tar.gz[gcp]
3. Initiate your Python script. To run your pipeline, use a command similar to the following example:

```
python my_pipeline.py --runner=DataflowRunner
--sdk_location=/path/to/apache-beam.tar.gz --project=my_project --region=us-central1
--temp_location=gs://my-bucket/temp ...
```

Note:

Install Beam from tarball on Dataflow worker takes minutes. One can also pass a prebuilt wheel to --sdk_location to save time

# Contributing back!

- New features
  - [XS] Useful configs but unexposed
  - [XL] New IO connectors
- Bug fixes
- Dependency upgrades
  - Resolves vulnerabilities from deps
  - Some IO pinned to very old client version
  - If newer dependency dropped Java8/11 support, still possible to upgrade ([#34858](#34858))

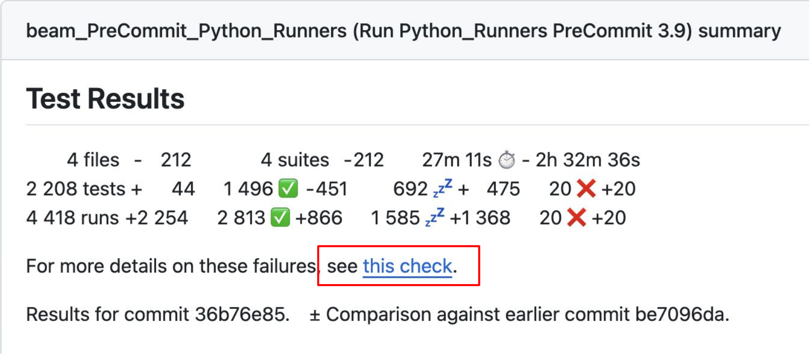
IBM Cloud
Pak for Data
An enterprise platform to operationalize data, analytics, and AI
Mara
Ruvalcaba
BEAM SUMMIT
Austin, TX
BEAM SUMMIT
Data Sorceress
SG WE<3DEVS
BEAM SUMMIT 2022 AUSTIN, TEXAS
aws
Google Open Source
Machine Learning Engineering with MLflow
Manage the end-to-end machine learning life cycle with

# Pull request

- A pull request will run PreCommit tests (GitHub workflow) automatically
- PreCommit tests can be retriggered by a Comment, e.g. ‘Run Java PreCommit’
- PostCommit tests can be manually triggered by touching a “trigger file” (under .github/tirgger_files)

Ref: https://github.com/apache/beam/blob/master/.github/workflows/README.md#running-workflows-manually

- Check test result
  - Go to workflow run -> Summary, it should show published test results



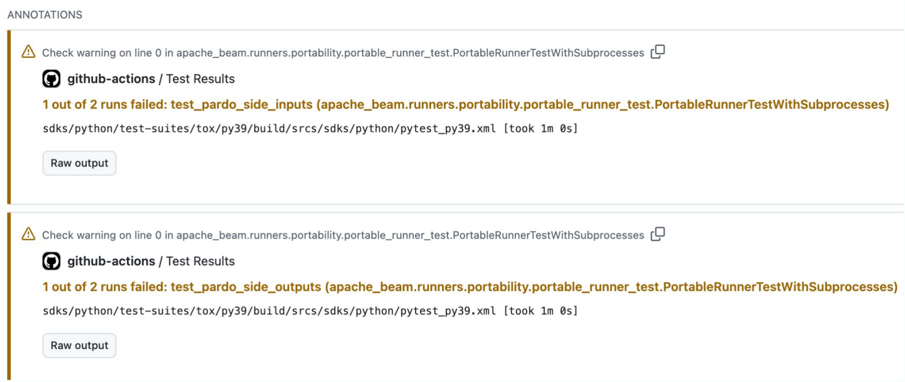
beam_PreCommit_Python_Runners (Run Python_Runners PreCommit 3.9) summary

## Test Results

```
    4 files  -    212        4 suites  -212      27m 11s ⏱ - 2h 32m 36s
2 208 tests +     44    1 496 ✅ -451      692 💤 +   475     20 ❌ +20
4 418 runs +2 254    2 813 ✅ +866    1 585 💤 +1 368     20 ❌ +20
```

For more details on these failures, see this check.

Results for commit 36b76e85.   ± Comparison against earlier commit be7096da.



ANNOTATIONS

⚠ Check warning on line 0 in apache_beam.runners.portability.portable_runner_test.PortableRunnerTestWithSubprocesses

github-actions / Test Results

1 out of 2 runs failed: test_pardo_side_inputs (apache_beam.runners.portability.portable_runner_test.PortableRunnerTestWithSubprocesses)

```
sdks/python/test-suites/tox/py39/build/srcs/sdks/python/pytest_py39.xml [took 1m 0s]
```

⚠ Check warning on line 0 in apache_beam.runners.portability.portable_runner_test.PortableRunnerTestWithSubprocesses

github-actions / Test Results

1 out of 2 runs failed: test_pardo_side_outputs (apache_beam.runners.portability.portable_runner_test.PortableRunnerTestWithSubprocesses)

```
sdks/python/test-suites/tox/py39/build/srcs/sdks/python/pytest_py39.xml [took 1m 0s]
```

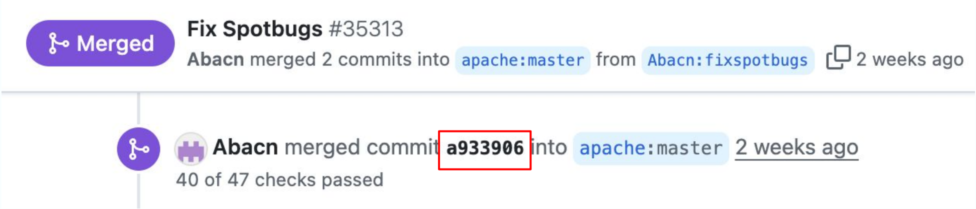
# Pull request: Reviewer

- A review bot automatically assign reviewers after tests passed
  - List of reviewers: https://github.com/apache/beam/blob/master/.github/REVIEWERS.yml
- Another are welcomed to add review comments, and can approve a PR. Then the bot will assign a commmitter for final approval and merge the PR

- We try to keep reviewer list up-to-date
  - If your institution made a contribution and/or regularly maintains a Beam components, consider add yourself to reviewer

# Check which release a PR in

- In merged PR

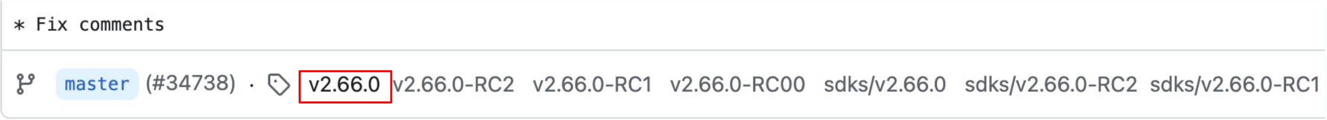


Merged **Fix Spotbugs** #35313

**Abacn** merged 2 commits into `apache:master` from `Abacn:fixspotbugs` 2 weeks ago

**Abacn** merged commit **a933906** into `apache:master` 2 weeks ago

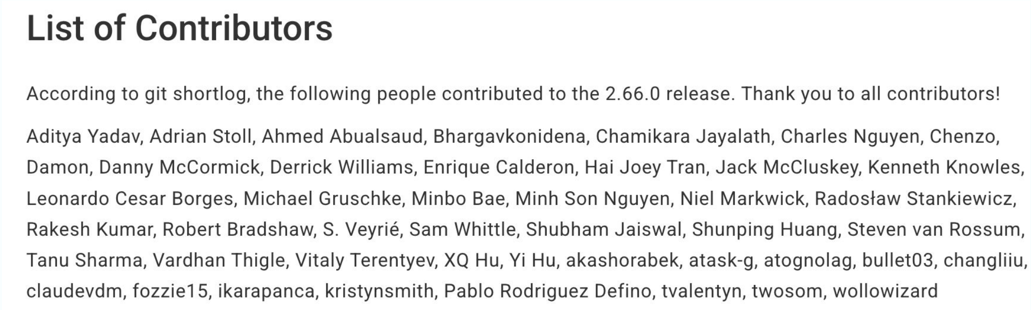
40 of 47 checks passed



```
* Fix comments
```

master (#34738) · v2.66.0 v2.66.0-RC2 v2.66.0-RC1 v2.66.0-RC00 sdks/v2.66.0 sdks/v2.66.0-RC2 sdks/v2.66.0-RC1

- Release blog: [https://beam.apache.org/blog/beam-2.66.0/](https://beam.apache.org/blog/beam-2.66.0/)



## List of Contributors

According to git shortlog, the following people contributed to the 2.66.0 release. Thank you to all contributors!

Aditya Yadav, Adrian Stoll, Ahmed Abualsaud, Bhargavkonidena, Chamikara Jayalath, Charles Nguyen, Chenzo, Damon, Danny McCormick, Derrick Williams, Enrique Calderon, Hai Joey Tran, Jack McCluskey, Kenneth Knowles, Leonardo Cesar Borges, Michael Gruschke, Minbo Bae, Minh Son Nguyen, Niel Markwick, Radosław Stankiewicz, Rakesh Kumar, Robert Bradshaw, S. Veyrié, Sam Whittle, Shubham Jaiswal, Shunping Huang, Steven van Rossum, Tanu Sharma, Vardhan Thigle, Vitaly Terentyev, XQ Hu, Yi Hu, akashorabek, atask-g, atognolag, bullet03, changliiu, claudevdm, fozzie15, ikarapanca, kristynsmith, Pablo Rodriguez Defino, tvalentyn, twosom, wollowizard

# Recap

- Beam monorepo structure
- Quick-start-java/python/go as minimum Templates user project
- Recent improvements for developer experiences
  - Documentations (code-change-guide.md)
  - Make docker optional for development (#33508)
  - Run examples in places on various runners (#35262)
  - IntelliJ integration (#35167)
  - Enable to compile Beam module that needs higher Java version than Java8 (#35167)

# Thank you!

Yi Hu

# QUESTIONS?

yhu@apache.org

Github: Abacn