

Growing the Apache Beam Community: Resources, Contributions, and Collaboration

Jana Polianskaja, Data Engineer @ Accenture Baltics

Danny McCormick, Software Engineer @ Google



Agenda

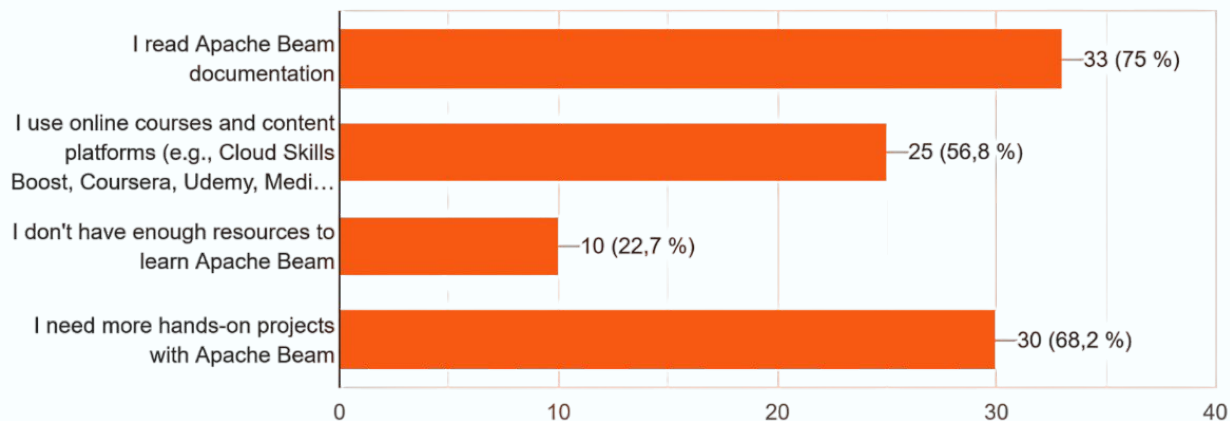


- Getting Started with Apache Beam
- Contribute to the Community!
- Code (or code-adjacent) Contributions
- Take Notes and Contribute Them Back
- Project Structure and Committership
- Google Summer of Code
- Beam Community: Future Plans & Initiatives
- Beam Community: Survey Comments

🔍 Getting Started with Apache Beam



What resources do you currently have available to learn and practice Apache Beam? Select all that apply.

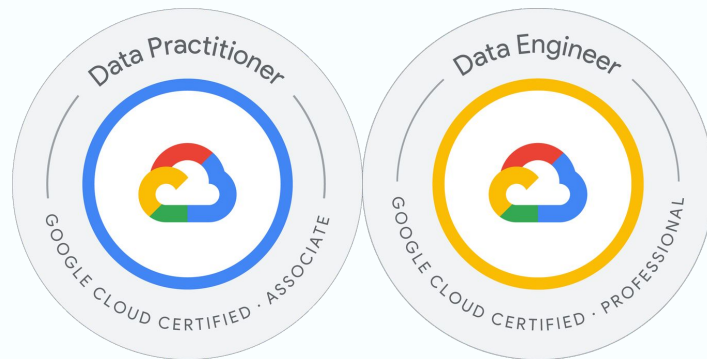


Apache Beam Community Survey Results

Getting Started with Apache Beam



<p>Google Cloud</p> <p>Serverless Data Processing with Dataflow: Foundations</p> <p>— ✓ —</p> <p>COMPLETION BADGE</p>	<p>Google Cloud</p> <p>Serverless Data Processing with Dataflow: Develop Pipelines</p> <p>— ✓ —</p> <p>COMPLETION BADGE</p>
<p>Google Cloud</p> <p>Data Warehousing for Partners: Process Data with Dataflow</p> <p>— ✓ —</p> <p>COMPLETION BADGE</p>	<p>Google Cloud</p> <p>Serverless Data Processing with Dataflow: Operations</p> <p>— ✓ —</p> <p>COMPLETION BADGE</p>
<p>Google Cloud</p> <p>Getting Started with Apache Beam</p> <p>— ✓ —</p> <p>COMPLETION BADGE</p>	<p>Google Cloud</p> <p>Dataflow Advanced</p> <p>— ✓ —</p> <p>INSIDE TRACK</p>



More info:

cloudskillsboost.google

<https://beam.apache.org/documentation/>

<https://github.com/GoogleCloudPlatform/training-data-analyst>

<https://github.com/apache/beam>

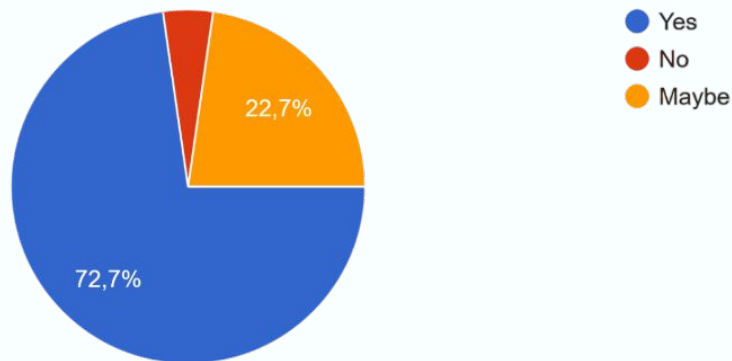
<https://cloud.google.com/products/dataflow>



Contribute to the Community!



Would you be interested in contributing to Apache Beam community (e.g., through pull requests, issue reporting, blog posts, etc.)?



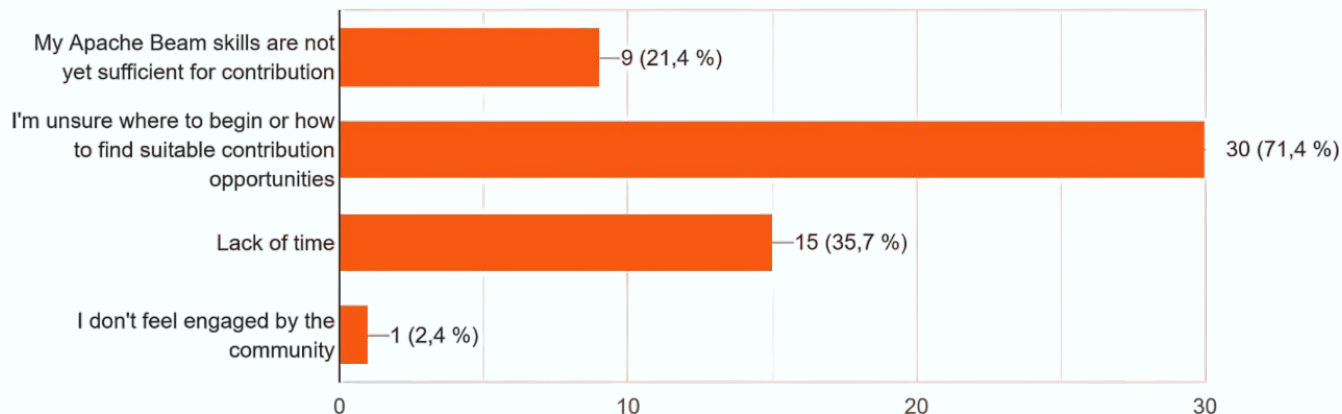
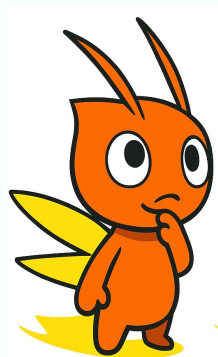
Apache Beam Community Survey Results



Contribute to the Community!



What are the biggest barriers preventing you from contributing to the Apache Beam community? Select all that apply.



Apache Beam Community Survey Results



Contribute to the Community!



Find an area to contribute

Beam recognizes code and non-code contributions.
Check out the [GitHub Issues](#) for some inspiration.



File a bug



Fix an open issue



Write a blogpost



Propose a feature



Share your success story



Produce a social media publication



Other ideas?

send a proposal to
dev@beam.apache.org

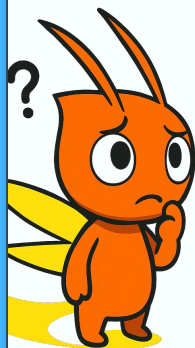


Improve documentation



Organize an event

<https://beam.apache.org/contribute/>





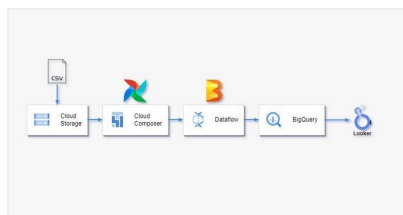
Contribute to the Community!



Google Cloud

Curated articles for developers

CONTAINERS / K8S DATA MACHINE LEARNING APP DEV INFRASTRUCTURE SECURITY / OPERATIONS ALL |



GCP Data Engineering Project: Building and Orchestrating an ETL Pipeline for Online Food...

This GCP Data Engineering project focuses on developing a robust ETL (Extract, Transform, Load) pipeline for the online food delivery...

jana_om
Dec 16 - 9 min read



Medium



ML/AI & GEN AI DATA CONTAINERS / K8S APP DEV INFRASTRUCTURE SECURITY / OPERATIONS ALL | MORE

TAGGED IN

Data



Google Cloud - Community

A collection of technical articles and blogs published or curated by Google Cloud Developer Advocates. The views expressed are those of the authors and don't necessarily reflect those of Google.

More information

FOLLOWERS

64K

ELSEWHERE

MORE, ON MEDIUM

Data



jana_om in Google Cloud - Community

May 20 - 4 min read



Beam College 2025: Online Training + Hackathon

medium.com/google-cloud/tagged/data



ML/AI & GEN AI DATA CONTAINERS / K8S APP DEV INFRASTRUCTURE SECURITY / OPERATIONS ALL | MORE

TAGGED IN

Data



Google Cloud - Community

A collection of technical articles and blogs published or curated by Google Cloud Developer Advocates. The views expressed are those of the authors and don't necessarily reflect those of Google.

More information

FOLLOWERS

54K

ELSEWHERE

MORE, ON MEDIUM

Data




jana_om in Google Cloud - Community

Apr 29 - 15 min read



GCP Data Engineering Project: Streaming Data Pipeline with Pub/Sub and Apache Beam/Dataflow

Read more...



July 23-25, 2024 @beamcollege.dev

Beam College 2024: Worth Your Time?

I recently attended Beam College for the second time, and I'm happy to share my thoughts. It's worth noting that this review comes from a...

jana_om
Jul 26 - 5 min read

medium.com/google-cloud/tagged/data



CONTAINERS / K8S DATA MACHINE LEARNING APP DEV INFRASTRUCTURE SECURITY / OPERATIONS ALL | MORE

TAGGED IN

Data



Google Cloud - Community

A collection of technical articles and blogs published or curated by Google Cloud Developer Advocates. The views expressed are those of the authors and don't necessarily reflect those of Google.

More information

FOLLOWERS

54K

ELSEWHERE

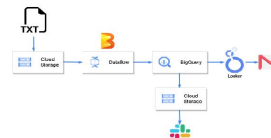
MORE, ON MEDIUM

Data



jana_om in Google Cloud - Community

Jan 31 - 14 min read



GCP Data Engineering Project: Connect Four game with Python and Apache Beam

Read more...

1



Contribute to the Community!



Google Cloud Innovators

THE OVERWHELMED PERSON'S GUIDE TO GOOGLE CLOUD | By [Richard Seroter](#)

[COMMUNITY CUTS]

- Can AI help you with your music playlist? If you give Sid 15 minutes, it looks like he can help. [This post](#) shows us how to use vector databases and Vertex AI for conducting a similarity search.
- Build Python APIs and deploy to Cloud Run. This is a [good post from Matt](#) who shows how to use the popular FastAPI framework for Python to build an API and deploy on a serverless stack.
- Let's see a serious data processing pipeline. Jana offers up a [detailed post](#) for setting up a streaming data pipeline and it teaches us more about stitching together a handful of cloud services.

PowerToGo Research Reports Categories Comparisons For Vendors [Sign In](#)

Home - Streaming Analytics | Google Cloud Dataflow

Google Cloud Dataflow Reviews

version 1.0.0.0

★★★★★ 4.0 out of 5 11 reviews 92% willing to recommend

Overview Reviews Pros & Cons Pricing Alternatives

Featured reviews

Jana Pollanskaja
Data Engineer at Accenture

★★★★★

As a data engineer, I find several features of Google Cloud Dataflow particularly valuable. The ability to test solutions locally using Direct Runner is crucial for development, allowing me to validate pipelines without incurring the costs of full Dataflow jobs. The unified programming...

[Read full review](#)

Amithaba Chakraborty
Senior Software Engineer at Google

★★★★★

The solution's setup process is a learning curve for us. It is a bit tricky overall. We receive vendor assistance in terms of infrastructure and expertise to build it. The time taken for deployment depends on the status of the environment, security and governance setup...

[Read full review](#)

Arpan Kashwaha
Senior Software Engineer at Google

★★★★★

We use Google Cloud Dataflow mainly for batch pipelines. We migrate workload from on-premise data movement to BigQuery or Storage Bucket. The most valuable features of Google Cloud Dataflow are scalability and connectivity. Google Cloud Dataflow should include a little...

[Read full review](#)

[Read more reviews](#)

beam About Get Started Documentation Roadmap Community Contribute Blog **Case Studies** [Search](#) [Feedback](#) [Apache](#)

Case Studies

Apache Beam powers many of today's leading projects, industry-specific use cases, and startups.

Accenture
Accenture Baltics' Journey with Apache Beam

Accenture Baltics uses Apache Beam on Google Cloud to build a robust data processing infrastructure for a sustainable energy leader. They use Beam to democratize data access, process data in real-time, and handle complex ETL tasks.

[Jana Pollanskaja](#)
Data Engineer @ Accenture Baltics

yelp
Building data abstractions with streaming at Yelp

At Yelp, Apache Beam allows teams to create custom streaming pipelines using Python, eliminating the need to switch to Scala or Java. This reduces the learning curve for Python developers and minimizes friction, while providing the flexibility to utilize existing Python libraries.

[Hakampreet Singh Pandher](#)
Software Engineer @ Yelp

in
Revolutionizing Real-Time Stream Processing: 4 Trillion Events Daily at LinkedIn

Apache Beam serves as the backbone of LinkedIn's streaming infrastructure, handling the near real-time processing of an astounding 4 trillion events daily through 3,000+ pipelines and thus powering personalized experiences for LinkedIn's vast network of over 950 million members

[Bingfeng Xia](#)
Engineering Manager @ LinkedIn

[Xinyu Liu](#)

← → ↺ [beam.apache.org/case-studies/](#) [Search](#) [Star](#) [Share](#) [Profile](#)

Confluence Spaces

Community Development

Pages / ComDev Wiki / GSoC

[\[GSOE\]\[Kwoks\] Support database backup to cloud storage](#)

- **Beam**
 - Enhancing Apache Beam JupyterLab Sidepanel for JupyterLab 4.x and Improved UI/UX
 - Beam ML Vector DB/Feature Store integrations
 - Simplify management of Beam infrastructure, access control and permissions via Platform features
 - Beam YAMLM, Iceberg, and Kafka User Accessibility
- **SkyWalking**
 - SkyWalking BanyanDB Extend remote.FS with Object Storage Support for AWS, Google Cloud, and Azure
- **Seata**
 - GSoC 2025 - Apache Seata(Incubating) Extend multi-raft cluster mode
 - GSoC 2025 - Apache Seata(Incubating)Unlocking the Power of Metadata in Apache Seata From Load Balancing to Advanced Routing
 - GSoC 2025 - Apache Seata(Incubating) Enhancing Connection Pool Management for Apache Seata AT/XA Transaction Modes
- **RocketMQ**
 - Refactoring the RocketMQ Dashboard UI and Enhancing Usability

[https://summerofcode.withgoogle.com/](#)

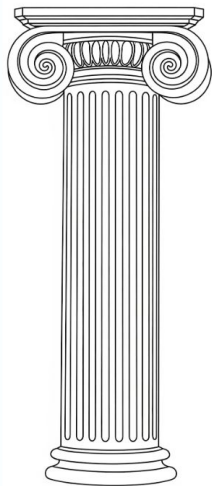


Code (or code-adjacent) Contributions

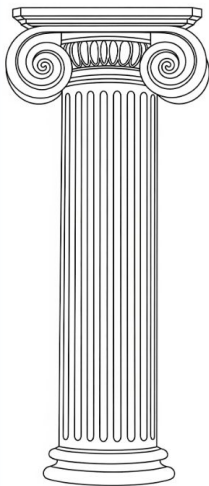


Find an area others are investing in

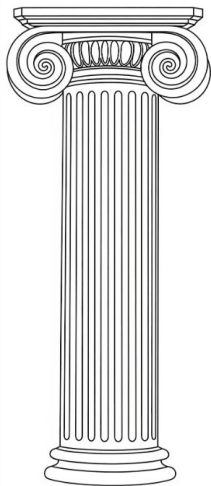
Access for All



Easy ops



First class ML



[2.64.0] - 2025-03-31

Highlights

- Managed API for [Java](#) and [Python](#) supports [key I/O connectors](#) Iceberg, Kafka, and BigQuery.

I/Os

- [Java] Use API compatible with both com.google.cloud.bigdataoss:util 2.x and 3.x in BatchLoads ([#34105](#))
- [IcebergIO] Added new CDC source for batch and streaming, available as `Managed.ICEBERG_CDC` ([#33504](#))
- [IcebergIO] Address edge case where bundle retry following a successful data commit results in data duplication ([#34264](#))
- [Java&Python] Add explicit schema support to JdbcIO read and xiang transform ([#23029](#))

New Features / Improvements

- [Python] Support custom coders in Reshuffle ([#29908](#), [#33356](#)).
- [Java] Upgrade SLF4J to 2.0.16. Update default Spark version to 3.5.0. ([#33574](#))
- [Java] Support for `--add-modules JVM` option is added through a new pipeline option `JdkAddRootModules`. This allows extending the module graph with optional modules such as SDK incubator modules. Sample usage: `<pipeline invocation> --jdkAddRootModules=jdk.incubator.vector` ([#30281](#)).
- Managed API for [Java](#) and [Python](#) supports [key I/O connectors](#) Iceberg, Kafka, and BigQuery.
- [YAML] Beam YAML UDFs (such as those used in MapToFields) can now have declared dependencies (e.g. pypl packages for Python, or extra jars for Java).
- Prism now supports event time triggers for most common cases. ([#31438](#))
 - Prism does not yet support triggered side inputs, or triggers on merging windows (such as session windows).

Breaking Changes

- [Python] Reshuffle now correctly respects user-specified type hints, fixing a previous bug where it might use FastPrimitivesCoder



[Feature Request]: Improve return/yield documentation

#34906 · damccorm opened on May 9

good first issue

new feature

P2

python

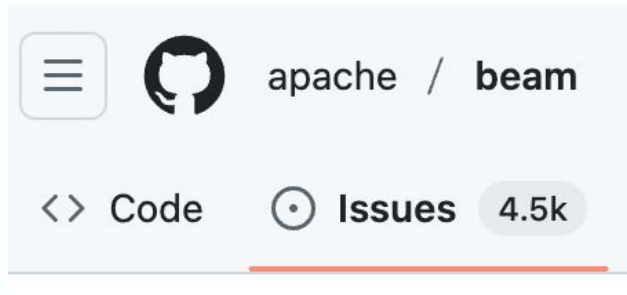




Code (or code-adjacent) Contributions



(Hard mode) Find an area you seem to care about more than anybody else





Take Notes and Contribute Them Back



Blog: My Experience at Beam College 2025 (3rd Place Hackathon Winner) #35315

Merged damccorm merged 4 commits into [apache:master](#) from [msugar:blog/anomaflow-hackathon-2025](#) 3 days ago

[Bug]: Worker logs on Prism Stand-Alone Binary Not Appearing #30



Listed in [#29650](#)



frankljin opened on Mar 21, 2024 · edited by frankljin

Edits ▾ ...

What happened?

To replicate this issue:

1. Clone the beam-starter-go pipeline [here](#)
2. Run it using the command. Worker logs should appear as expected ("Hello", "World!", "Greetings", one per line)

[VOTE] Release 2.65.0, release candidate #1

Thoughts from a first time contributor

Posted to [dev@beam.apache.org](#)



Danny McCormick - Monday, January 31, 2022 3:17:03 PM EST

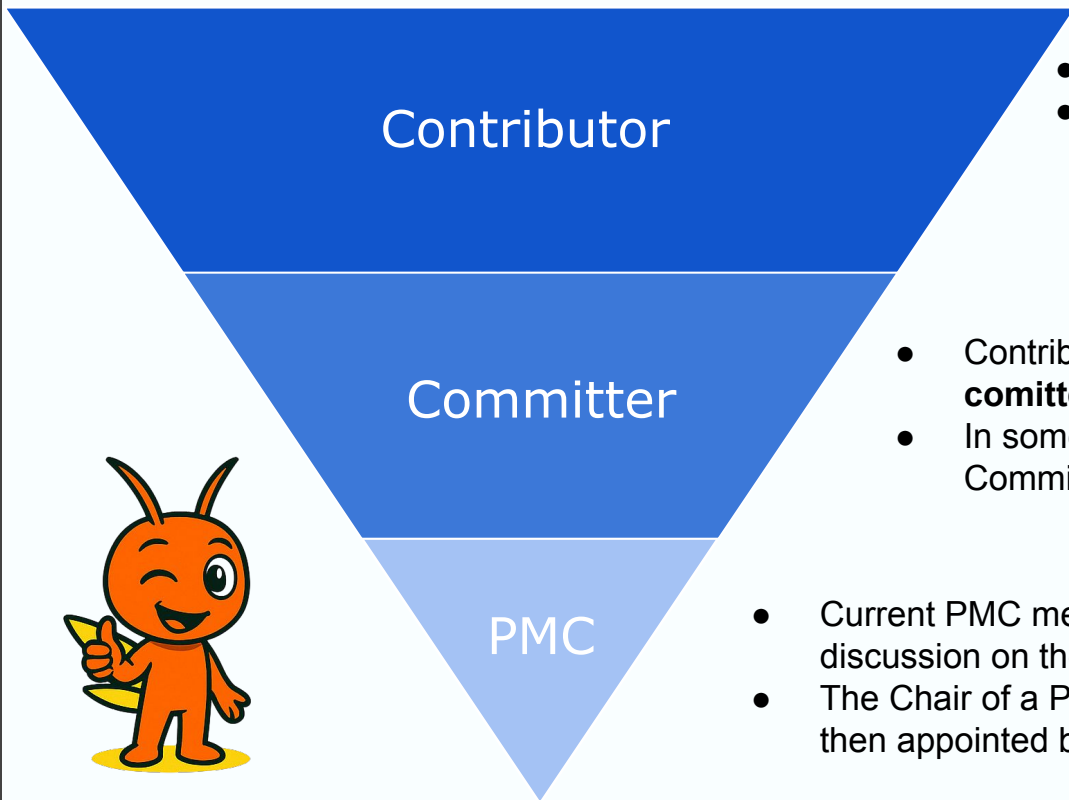
👋 Hey folks, my name is Danny - I recently completed my first Beam PR[1] (a small extension to the Go Dataflow runner) and am planning on becoming a more regular part of the community. As such, I wanted to use my fresh newbie eyes and share some of what was nice and where there was friction about getting started.

Disclaimer: this is coming from the perspective of someone who is pretty used to open source development, but has minimal experience with the Apache way, Beam, and the languages my change came in. I'm hoping my experience is helpful to those of you who have been around for a while and haven't seen things as a newcomer in a long time, but it may not be reflective of the experience of others.

Things that were really nice:



Project Structure and Committership



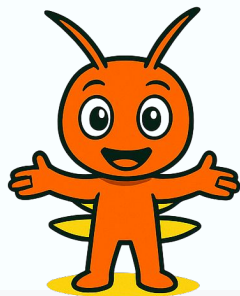
- Anyone can be a contributor, **not just coders**
- All **CoPDoC** contributions earn merit: Community, Project vision, Documentation and Code

- Contributors are observed by the PMC to be **awarded comittership according to their merit**
- In some projects (more common in incubating projects) Committees = PMC

- Current PMC member **nominates** an individual through discussion on the private@ mailing list
- The Chair of a PMC is selected by PMC members and then appointed by the Board.



Google Summer of Code



Apply

Interested contributors propose a project to work on.



Code

Accepted GSoC contributors spend the summer coding with guidance from a mentor.



Share

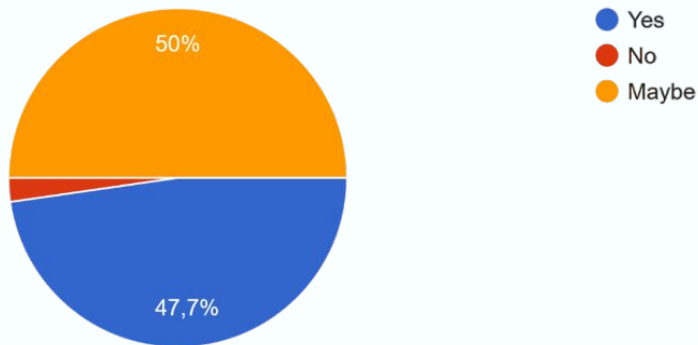
Submit your code for the world to use!



Beam Community: Future Plans & Initiatives



Would you be interested in taking an Apache Beam certification exam?



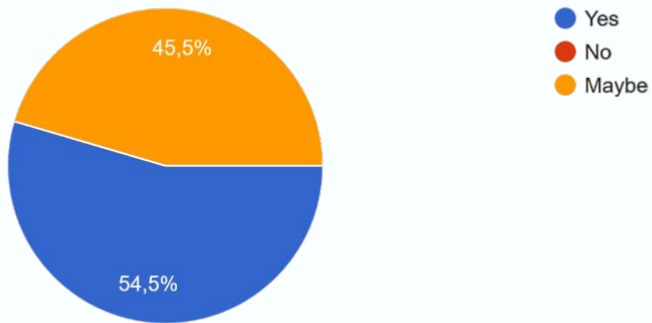
Apache Beam Community Survey Results




Beam Community: Future Plans & Initiatives



Would you attend future online/in-person Apache Beam meetups?





Belgium Apache Beam Meetup

3.7 ★★★★★ 7 ratings

Brussels, Belgium


138 members · Public group

Organized by Alex Van Boxel and 1 other

Share: [Facebook] [Twitter] [LinkedIn] [Email]

About Events Members Photos Discussions

[Join this group](#)



Zürich Apache Beam Meetup

4.7 ★★★★★ 12 ratings

Zürich, Switzerland

88 members · Public group

Organized by Tectrain

Share: [Facebook] [Twitter] [LinkedIn] [Email]

About Events Members Photos Discussions

[Join this group](#)

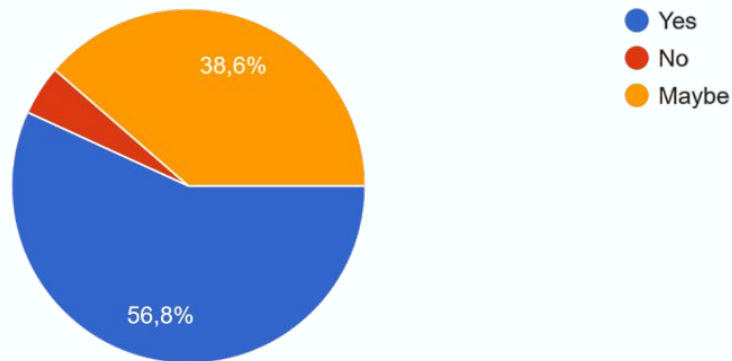
Apache Beam Community Survey Results



Beam Community: Future Plans & Initiatives



Would you be interested in participating in Apache Beam Community Ambassador program?



Apache Beam Community Survey Results



Do you have any suggestions for improving the Apache Beam community?

- Pair new contributors with experienced mentors for a period of time.
- Make it more beginner friendly, specially to get more users from Spark community
- I hope Beam Summit will be held in Asia as well.
- How can we become more involved???



Thank you!

QUESTIONS?

Jana Polianskaja
Medium: [@jana_om](#)
Linkedin: [jana-polianskaja](#)
Github: [janaom](#)

Danny McCormick
Linkedin: [danny-mccormick](#)
Github: [damccorm](#)