

Apache Beam version 2.5.0

We are glad to present the new 2.5.0 release of Beam. This release includes multiple fixes and new functionalities. For more information, please, check the detailed release notes.

New Features / Improvements

Go SDK support

The Go SDK has been officially accepted into the project, after an incubation period and community effort. Go pipelines run on Dataflow and all portable runners, notably the Flink runner. More details are [here](https://beam.apache.org/documentation/sdks/go/).

Parquet support

Support for Apache Parquet format was added. It uses Parquet 1.10 release which, thanks to AvroParquetWriter's API changes, allows FileIO.Sink implementation.

Performance/Integration Tests

- Added new integration tests - HCatalogIOIT (Hive), HBaseIOIT, ParquetIOIT (with the IO itself, local filesystem, HDFS)
- Multinode (3 data node) HDFS cluster is used for running tests on HDFS.
- Several improvements on performance tests running and results analysis.
- Scaled up Kubernetes cluster from 1 to 3 nodes.
- Added metrics in Spark streaming.

Internal Build System: Migrated to Gradle

After a months-long community effort, the internal Beam build has been migrated from Maven to Gradle. The new build system was chosen because of dependency-driven build support, incremental build/test, and support for non-Java languages.

Nexmark Improvements

- Kafka support as a source/sink for events and results.

- Translation of all queries to Beam SQL.

Beam SQL

- Support for MAP, ROW, ARRAY data types
- Support UNNEST on array fields
- Improved optimizations
- Upgrade Calcite to 1.16
- Support SQL on POJOs via automatic conversion
- Schema moved into core Beam
- UDAFs can be indirect subclasses of CombineFn
- *Many* other small bugfixes

Schema-Aware PCollections

TODO

Portability

- Common shared code related to supporting portable execution for runners
- Python SDK supporting side inputs over the portability APIs

Extract metrics in a runner agnostic way

Metrics are pushed by the runners to configurable sinks (Http REST sink available). It is already enabled in Flink and Spark runner, work is in progress for Dataflow.

Miscellaneous Fixes

SDKs

- Implemented HDFS FileSystem for Python SDK.
- Python SDK adds support for side inputs for streaming execution.

Runners

- Updated Spark runner to Spark version 2.3.1
- Fixed issue with late elements windowed into expired fixed windows get dropped in Directrunner.

IOs

- CassandraIO gained a better split algorithm based on overlapping regions.
- ElasticsearchIO supports partial updates and allows to pass id, type and index per document.
- SolrIO supports a more robust retry on write strategy.
- S3 fileSystem support encryption (SSE-S3, SSE-C and SSE-KMS).
- Improved connection management in JdbcIO.
- Added support the element timestamps while publishing to Kafka.

Other

- Using Java ErrorProne for static analysis.

List of Contributors

According to git shortlog, the following 84 people contributed to the 2.5.0 release. Thank you to all contributors!

Ahmet Altay, Alan Myrvold, Alex Amato, Alex Van Boxel, Alexander Dejanovski, Alexey Romanenko, Aljoscha Krettek, ananvay, Andreas Ehrencrona, Andrew Pilloud, Ankur Goenka, Anton Kedin, arkash, Austin Bennett, Axel Magnuson, Ben Chambers, Ben Sidhom, Bill Neubauer, Boyuan Zhang, Braden Bassingthwaite, Cade Markegard, cclauss, Chamikara Jayalath, Charles Chen, Chuan Yu Foo, Cody Schroeder, Colm O hEigartaigh, Daniel Oliveira, Dariusz Aniszewski, David Cavazos, Dawid Wysakowicz, Eric Roshan-Eisner, Etienne Chauchot, Eugene Kirpichov, Flavio Fiszman, Geet Kumar, GlennAmmons, Grzegorz Kołakowski, Henning Rohde, Innocent Djiofack, Ismaël Mejía, Jack Hsueh, Jason Kuster, Javier Antonio Gonzalez Trejo, Jean-Baptiste Onofré, Kai Jiang, Kamil Szewczyk, Katarzyna Kucharczyk, Kenneth Jung, Kenneth Knowles, Kevin Peterson, Lukasz Cwik, Łukasz Gajowy, Mairbek Khadikov, Manu Zhang, Maria Garcia Herrero, Marian Dvorsky, Mark Liu, Matthias

Feys, Matthias Wessendorf, mingmxu, Nathan Howell, Pablo Estrada, Paul Gerver, Raghu Angadi, rarokni, Reuven Lax, Rezan Achmad, Robbe Sneyders, Robert Bradshaw, Robert Burke, Romain Manni-Bucau, Sam Waggoner, Sam Whittle, Scott Wegner, Stephan Hoyer, Thomas Groh, Thomas Weise, Tim Robertson, Udi Meiri, Valentyn Tymofieiev, XuMingmin, Yifan Zou, Yunqing Zhou