Beam Streaming IO Ecosystem

Evolution towards production readiness

Yi Hu
Google
GitHub: @Abacn

BEAM SUMMIT
NYC 2026

Agenda

- Beam IO overview
- Streaming IO
  - SDK interfaces (UnboundedSource, SDF)
  - From “M0” existence to “M1” resilience and beyond
- Recent streaming IO improvments
  - Debezium, JmsIO, MqttIO, PulsarIO
- Contributing to Beam is easier than ever
  - AI tooling

Apache Beam IO connectors

https://beam.apache.org/documentation/io/connectors/

- 65 listed built-in IO (under apache/beam repo)
- 11 listed "other" IO (third-party)

- **Code velocity metrics (#merged PR)**

IO Name	PRs (1 Year)	PRs (3 Years)
Gcp-bigquery	60	210
Python-Gcp	51	186
Iceberg	41	95
Kafka	40	111
Gcp-spanner	38	86
Jdbc	22	50
Gcp-bigtable	11	55
AWS2	10	27
Debezium	10	17
Gcp-pubsub**	9	37
Rrio	7	27
Hbase	7	8
Gcp-firestore	6	15
Elasticsearch	6	12
Gcp-healthcare	6	11
Solace	5	27
Jms	5	12
Snowflake	5	11
Mongodb	5	10
Pulsar	5	7
Clickhouse	4	11
Synthetic	4	6
Hadoop-format	4	5
Datadog	4	4
Cassandra	3	7
Splunk	3	6
Parquet	3	5
RabbitMQ	3	5
DeltaLake	3	3
Gcp-datastore	2	10
Python-Azure	2	10
Thrift	2	7
Mqtt	2	6

Apache Beam IO connectors

https://beam.apache.org/documentation/io/connectors/

- 65 listed built-in IO (aka under apache/beam)
- 11 listed "other" IO (third-party)

- **IO with streaming source support**

IO Name	PRs (1 Year)	PRs (3 Years)
Gcp-bigquery	60	210
Python-Gcp	51	186
Iceberg	41	95
Kafka	40	111
Gcp-spanner	38	86
Jdbc	22	50
Gcp-bigtable	11	55
AWS2	10	27
Debezium	10	17
Gcp-pubsub**	9	37
Rrio	7	27
Hbase	7	8
Gcp-firestore	6	15
Elasticsearch	6	12
Gcp-healthcare	6	11
Solace	5	27
Jms	5	12
Snowflake	5	11
Mongodb	5	10
Pulsar	5	7
Clickhouse	4	11
Synthetic	4	6
Hadoop-format	4	5
Datadog	4	4
Cassandra	3	7
Splunk	3	6
Parquet	3	5
RabbitMQ	3	5
DeltaLake	3	3
Gcp-datastore	2	10
Python-Azure	2	10
Thrift	2	7
Mqtt	2	6

Streaming sources

Unbounded Source: the classic API for reading streaming data.

- Enables processing of infinite data streams with checkpointing and watermarks.
- Provides a robust mechanism for fault-tolerant data ingestion in real-time pipelines.

Splittable DoFn (SDF): unified API for building source connectors.

- Enables fine-grained parallelization by splitting processing of a single element across multiple workers.
- Unifies the programming model for both bounded (batch) and unbounded (streaming) data sources.

UnboundedSource & UnboundedReader

- Milestone 0 (minimal requirements)
  - UnboundedSource.createReader()
  - UnboundedReader.start()
  - .getCurrent()
  - .getCurrentTimestamp()

- Milestone 1 (data integrity requirements)
  - UnboundedSource.getCheckpointMark() **Checkpointing**
  - CheckpointMark.finalizeCheckpoint()
  - UnboundedReader.getWatermark() **Watermark management**
  - UnboundedReader.getCurrentRecordId() **Deduplication**

- Milestone 2+ (scalability, performance)
  - UnboundedSource.split() **Sharding**
  - Benchmarking

Splittable DoFn

- Milestone 0 (minimal requirements)
  - @ProcessElement
  - RestrictionTracker.tryClaim()
  - RestrictionTracker.currentRestriction()

- Milestone 1 (data integrity requirements)
  - RestrictionTracker.tryClaim() and checkDone() **Checkpointing**
  - @GetInitialRestriction **pipeline restart**
  - @TruncateRestriction **pipeline drain**
  - RestrictionTracker.getWatermark() **Watermark management**
  - **Runner deduplication not supported

- Milestone 2+ (scalability, performance)
  - RestrictionTracker.trySplit() **Sharding**
  - @GetSize / @GetProgress
  - Benchmarking

DebeziumIO: Recent improvements

A platform for Change Data Capture (CDC). It monitors databases—such as PostgreSQL, MySQL, MongoDB, and Oracle via native transaction logs

- #34747 Upgrade (from 1.x) to 3.1
  - requires Java17+, while Beam was on Java8 (Java11 currently)
- #28248 Pipeline restart from offset
  - An "OffsetRetainer" framework save offset on-the-fly and load on pipeline restart

Community contributors: tkaymak@ jiufeng-liu-ck@

JmsIO: Recent improvements

Java Message Service (JMS): a standard and classical Java API for messaging, widely used by enterprises

- #30054 JmsIO checkpointMark immutable
  - fixes dataloss on Dataflow autoscaling
- #30253 Implementing requiresDeduping
  - Support runner deduplication (Dataflow streaming feature)

Challenge: inconsistent behavior in different Jms providers beyond Jms spec

- #32483 NPE seen after #30253 (fixed)
- #30218 (Comment) unacked knowledges in some clients due to caching and unfinalized checkpoint

A whole new SolaceIO was introduced: https://beamsummit.org/sessions/2024/solace/

MqttIO: recent improvements

MQTT (Message Queuing Telemetry Transport) a lightweight, publish-subscribe messaging protocol primarily used for Internet of Things (IoT) and machine-to-machine (M2M) communication.

- #19376 Support Dynamic destinations in write
- #32195 Support Read with Metadata
- #36053 Fix checkpointMark
  - previously wrong implementation: checkpointMark shared across checkpoints

Ongoing:

- #21060 Python Schema Transform support

Community contributors: tkaymak@ twosom@

PulsarIO: recent improvements

Apache Pulsar, a distributed pub-sub messaging and event-streaming platform.

Beam's PulsarIO was in incomplete status (#31078) javadoc not published and not shown in website.

After #36141 (Beam 2.69.0), PulsarIO (read, write) is functioning (i.e. completed M0).

Open tasks:

- Implement consumer based read

Contributing to Beam!

- With AI tooling, contributing to Beam is easier than ever
- AI is good at (even without dedicate prompt or nested SKILLs):
  - Coding following existing pattern (UnboundedSource; SDF; SchemaTransform)
  - Backporting functionalities (Java -> Python)

- As **implementing** becomes easier, **identifying task/issue** serves more important roles in terms of evolution of a community supported IO connector

Contributing to IO Ecosystem

- "good first issue" label
  - small tasks: new feature; feature gap; test coverage; etc
- Google Summer of Code (GSoC)
  - Concrete but well-defined projects
  - This year: Porting UnboundedSource (#19137) and Watch transform (#21521) to Beam Python SDK
- Report GitHub Issues!
- Share your user journey, blog posts

Yi Hu

QUESTIONS?

yhu@apache.org

Github: Abacn

BEAM SUMMIT NYC 2026