Beam YAML and Protobuf

Ferran Fernandez
Austin Bennett

BEAM SUMMIT

September 4-5, 2024

Sunnyvale, CA. USA

About us

Austin Bennett

Ferran Fernandez

Chartboost

- ADS!

- "Build your mobile business with the leading in-app monetization and programmatic advertising platform"

Chartboost

© Chartboost. All rights reserved.

Agenda

- YAML, Protobuf & Beam
- Why Beam YAML?
- Beam YAML use case
- Findings & Limitations
- Conclusion & Takeaways
- Q&A

Chartboost

© Chartboost. All rights reserved.

YAML, Protobuf & Beam

© Chartboost. All rights reserved.

Why YAML

- Prevalent across industry
  - Esp. as config
- LOL – “No Code”

- Aside →
  - Pkl is emerging as interesting/related
  - See: https://pkl-lang.org/index.html

Chartboost

© Chartboost. All rights reserved.

Why Use PROTO

- https://protobuf.dev/
- Data Types
- Structured
- LOTs of use cases
  - Also see gRPC
- Some efficiencies vs alternatives
  - Naturally pros/cons

protobuf

Protocol Buffers

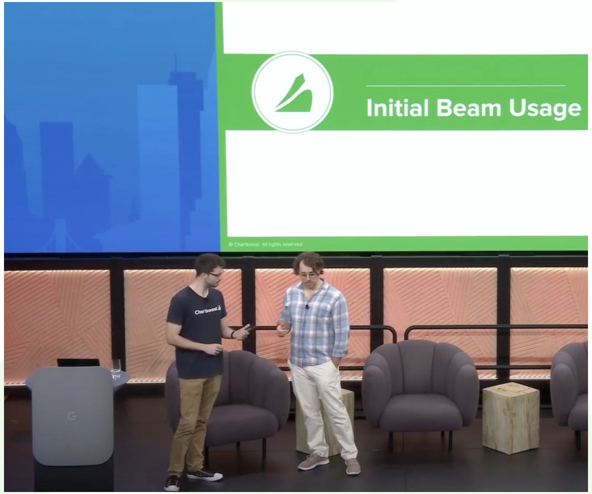
Chartboost

© Chartboost. All rights reserved.

Beam YAML + Proto

These now work well together!

Proto is just one way of representing the data.



Some of the background can be found
 in our talk at Beam Summit 2023.

Why Beam YAML?

© Chartboost. All rights reserved.

Current challenges

These are the main challenges we've seen at Chartboost:

- Complex pipeline setup:
  - It may take some time for a newcomer engineer without experience with Beam to set up their first pipeline.
  - Download all the dependencies, debug, test, etc.

- High maintenance and operational costs:
  - Once the pipeline is running in production, you must maintain and upgrade it.

- Limited reusability and scalability
  - Some custom implementations could lack flexibility, making it challenging to reuse.

Chartboost

© Chartboost. All rights reserved.

Solution

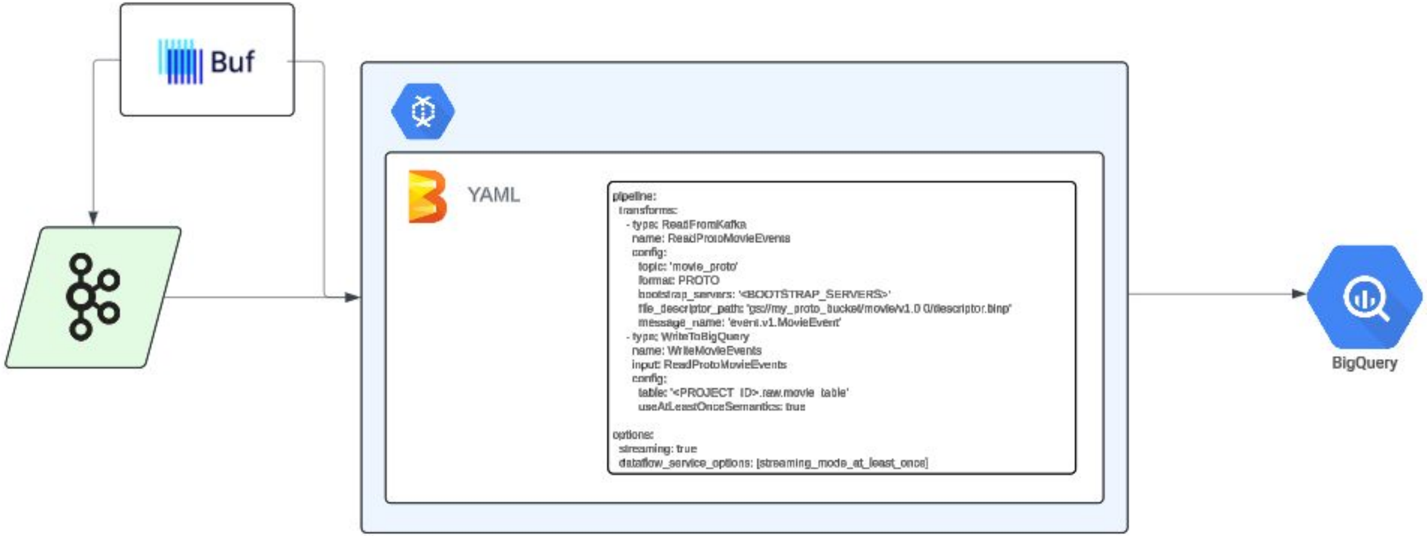
- Reusability
- No-code development (*)
- Extensibility
- Declarative Language
- Backwards Compatibility

Chartboost

© Chartboost. All rights reserved.

Beam YAML use case

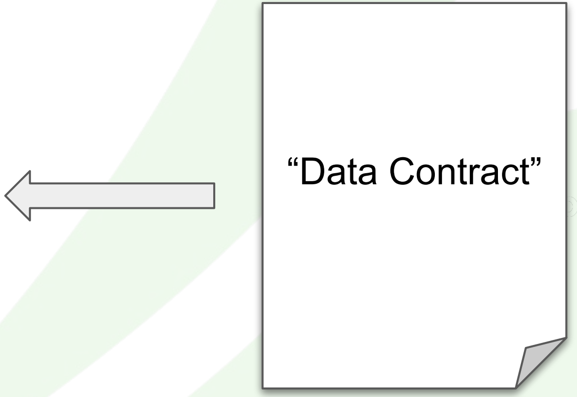
Architecture Diagram



```
pipeline:
  transforms:
    - type: ReadFromKafka
      name: ReadProtoMovieEvents
      config:
        topic: 'movie_proto'
        format: PROTO
        bootstrap_servers: '<BOOTSTRAP_SERVERS>'
        file_descriptor_path: 'gs://my_proto_bucket/movie/v1.0.0/descriptor.binp'
        message_name: 'event.v1.MovieEvent'
    - type: WriteToBigQuery
      name: WriteMovieEvents
      input: ReadProtoMovieEvents
      config:
        table: '<PROJECT_ID>.raw.movie_table'
        useAtLeastOnceSemantics: true

options:
  streaming: true
  dataflow_service_options: [streaming_mode_at_least_once]
```

Let’s start with our events

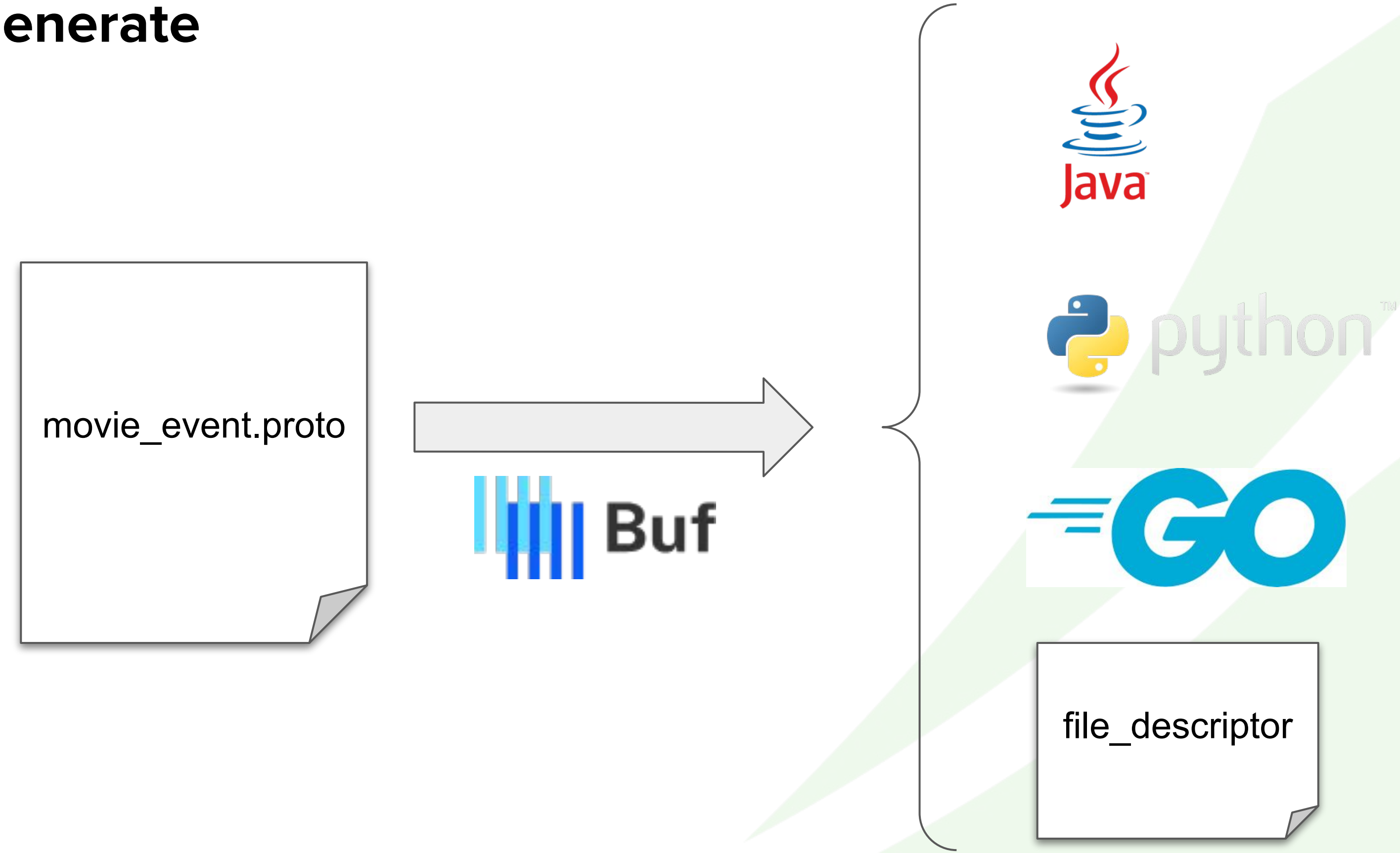
```
syntax = "proto3";

package event.v1;

import "bq_field.proto";
import "bq_table.proto";
import "buf/validate/validate.proto";
import "google/protobuf/wrappers.proto";

message MovieEvent {
  option (gen_bq_schema.bigquery_opts).table_name = "movie_table";
  google.protobuf.StringValue event_id = 1 [(gen_bq_schema.bigquery).description = "Unique Event ID"];
  google.protobuf.StringValue user_id = 2 [(gen_bq_schema.bigquery).description = "Unique User ID"];
  google.protobuf.StringValue movie_id = 3 [(gen_bq_schema.bigquery).description = "Unique Movie ID"];
  google.protobuf.Int32Value rating = 4 [(buf.validate.field).int32 = {
    // validates the average rating is at least 0
    gte: 0,
    // validates the average rating is at most 100
    lte: 100
  }, (gen_bq_schema.bigquery).description = "Movie rating"];
  string event_dt = 5 [
    (gen_bq_schema.bigquery).type_override = "DATETIME",
    (gen_bq_schema.bigquery).description = "UTC Datetime representing when we received this event. Format: YYYY-MM-DDTHH:MM:SS",
    (buf.validate.field) = {
      string: {
        pattern: "^\\d{4}-\\d{2}-\\d{2}T\\d{2}:\\d{2}:\\d{2}$"
      },
      ignore_empty: false,
    }
  ];
}
```



“Data Contract”

© Chartboost. All rights reserved.

Buf generate



movie_event.proto

file_descriptor

Chartboost

© Chartboost. All rights reserved.

Beam YAML Configuration

```
pipeline:
  transforms:
    - type: ReadFromKafka
      name: ReadProtoMovieEvents
      config:
        topic: 'movie_proto'
        format: PROTO
        bootstrap_servers: '<BOOTSTRAP_SERVERS>'
        file_descriptor_path: 'gs://my_proto_bucket/movie/v1.0.0/descriptor.binp'
        message_name: 'event.v1.MovieEvent'
    - type: WriteToBigQuery
      name: WriteMovieEvents
      input: ReadProtoMovieEvents
      config:
        table: '<PROJECT_ID>.raw.movie_table'
        useAtLeastOnceSemantics: true

options:
  streaming: true
  dataflow_service_options: [streaming_mode_at_least_once]
```

Chartboost

© Chartboost. All rights reserved.

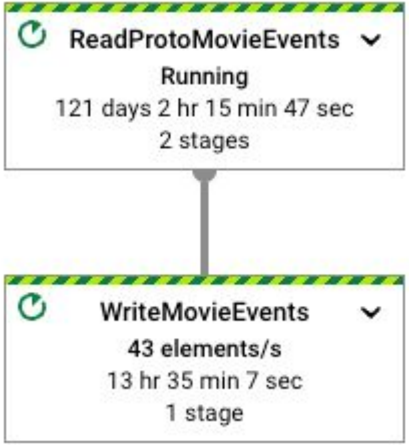
Terraform deployment

```
resource "google_dataflow_flex_template_job" "data_movie_job" {
  provider                  = google-beta
  project                  = var.gcp_project_id
  name                      = "movie-proto-events"
  container_spec_gcs_path      =
"gs://dataflow-templates-${var.gcp_region}/2024-05-13-00_RC00/flex/Yaml_Template"
  region                   = var.gcp_region
  on_delete                 = "drain"
  machine_type                = "n2d-standard-4"
  enable_streaming_engine      = true
  subnetwork                  = var.subnetwork
  skip_wait_on_job_termination = true
  parameters = {
    yaml_pipeline_file =
"gs://${var.bucket_name}/yamls/${var.package_version}/movie_events_pipeline.yml"
    max_num_workers    = 40
    worker_zone        = var.gcp_zone
  }
  depends_on = [google_project_service.enable_dataflow_api]
}
```

Chartboost

© Chartboost. All rights reserved.

Result



ReadProtoMovieEvents

Running

121 days 2 hr 15 min 47 sec

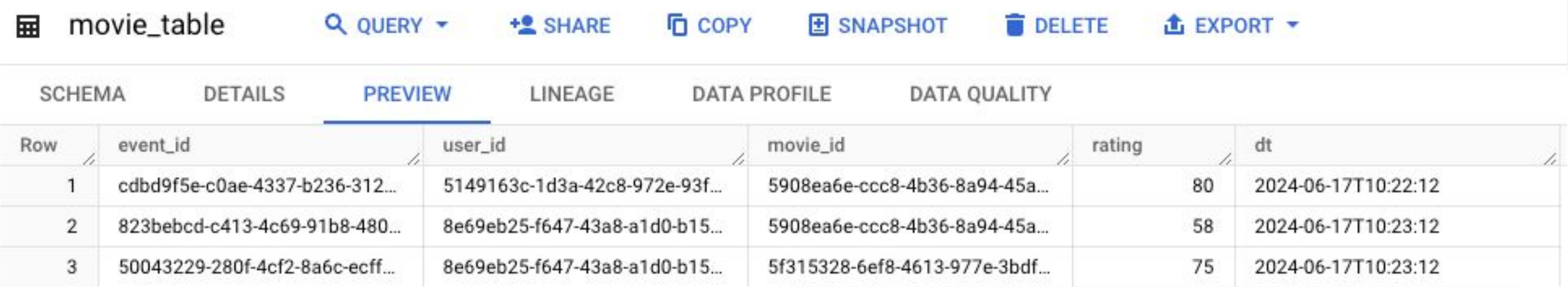
2 stages

WriteMovieEvents

43 elements/s

13 hr 35 min 7 sec

1 stage



Row	event_id	user_id	movie_id	rating	dt
1	cdbd9f5e-c0ae-4337-b236-312...	5149163c-1d3a-42c8-972e-93f...	5908ea6e-ccc8-4b36-8a94-45a...	80	2024-06-17T10:22:12
2	823bebcd-c413-4c69-91b8-480...	8e69eb25-f647-43a8-a1d0-b15...	5908ea6e-ccc8-4b36-8a94-45a...	58	2024-06-17T10:23:12
3	50043229-280f-4cf2-8a6c-ecff...	8e69eb25-f647-43a8-a1d0-b15...	5f315328-6ef8-4613-977e-3bdf...	75	2024-06-17T10:23:12

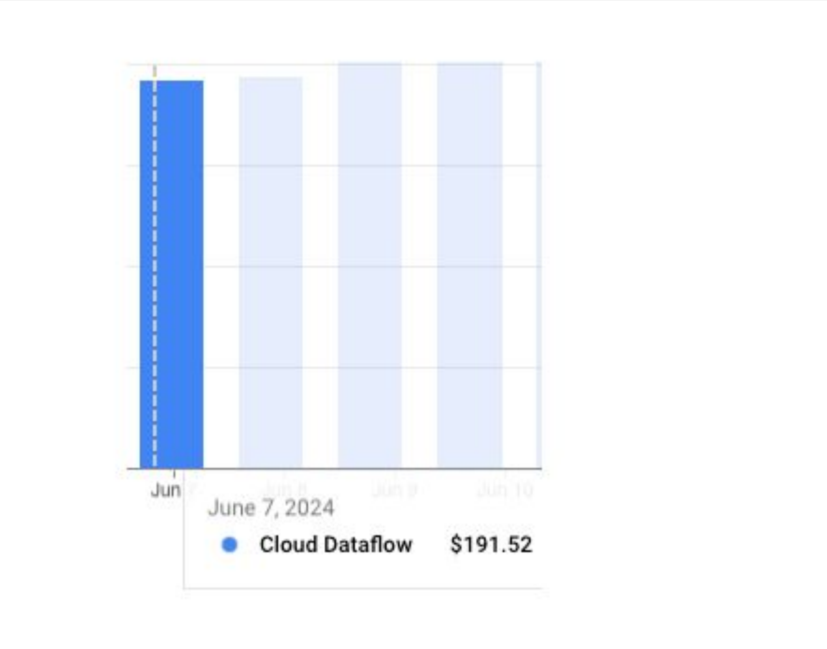
Chartboost

© Chartboost. All rights reserved.

Findings & Limitations

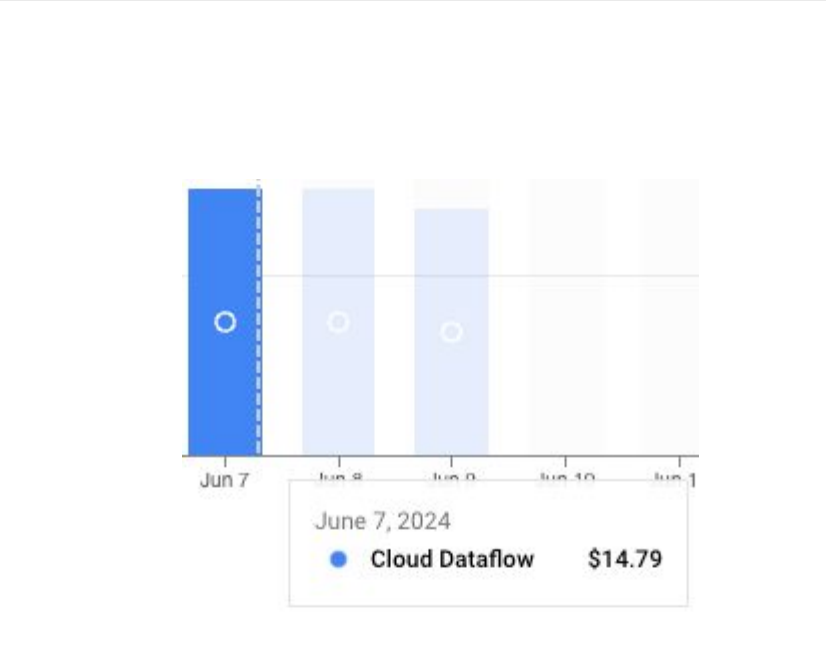
© Chartboost. All rights reserved.

Findings: Protobuf cost efficiency



June 7, 2024

● Cloud Dataflow $191.52



June 7, 2024

● Cloud Dataflow $14.79

JSON Events:

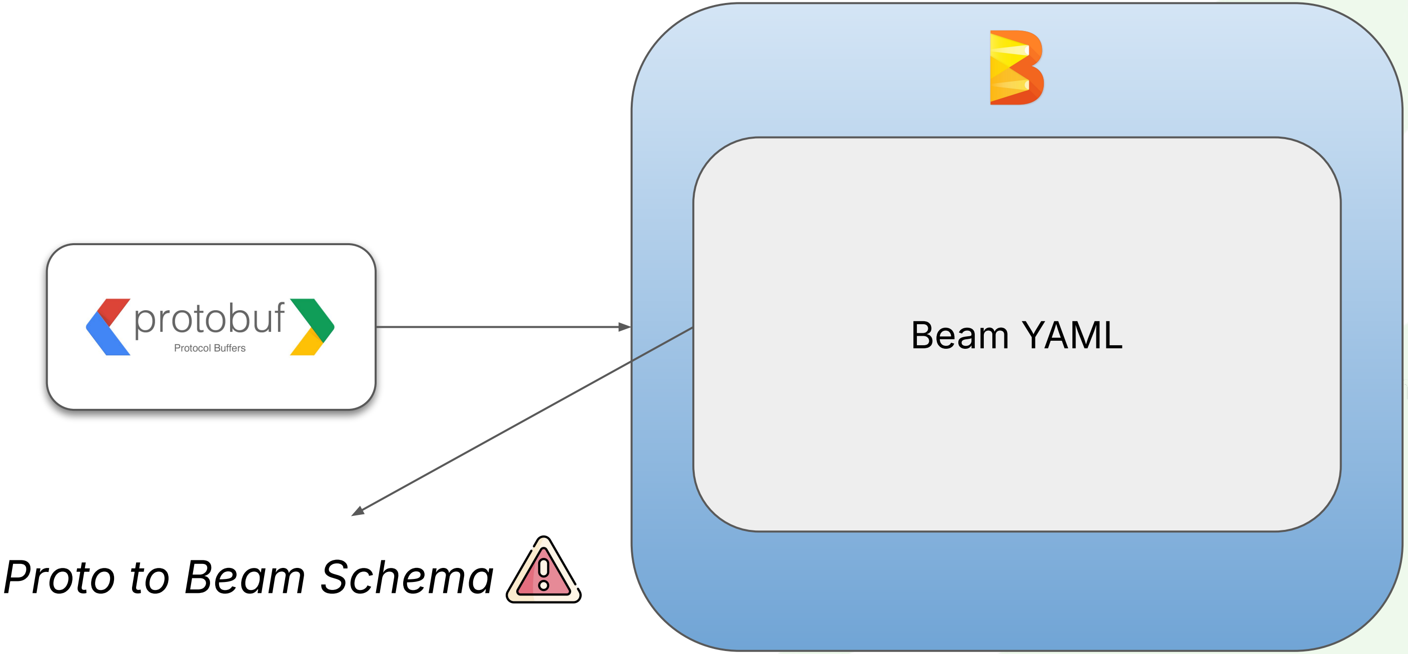
- 3500 events / s (each)
- 5 pipelines

Proto Events:

- 3500 events / s (each)
- 5 pipelines

© Chartboost. All rights reserved.

Limitations: Protobuf to Beam Schema



Proto to Beam Schema ⚠

Limitations: Still some missing features

- Beam YAML still don't support all I/Os.

- KafkaIO in Beam YAML only supports the Confluent Schema Registry. Ideally, we could extend it to support multiple schema registries. (Buf, Apicurio, etc.)

- Documentation has improved, but it could be better, perhaps by including more transformations and multilingual examples. This is where we encourage the community to jump in and help with this. https://s.apache.org/beam-yaml-contribute

Chartboost

© Chartboost. All rights reserved.

Conclusions & Takeaways

© Chartboost. All rights reserved.

Conclusion and Takeaways

- Beam YAML has a lot of positives.
  - **Low Learning Curve:** Beam YAML is easy to learn, enabling teams to get up to speed quickly.
  - **Faster Iterations:** The simplicity of Beam YAML allows for faster and more efficient iterations.
  - **Proto Introduction:** The integration of Proto supports the shift-left philosophy, enabling teams to "fail early and fix quickly."
    - Besides that, it also lowered processing costs due to the efficiency of Proto.

Chartboost

© Chartboost. All rights reserved.

Thank you!

Questions?

You can reach out via Linkedin:

Email: fferngar@proton.me

Austin: austin@apache.org

Some notes for extending IOs

- This is an Open Source Project!! :-)

- https://github.com/apache/beam/blob/master/sdks/python/apache_beam/io/kafka.py#L115
- Ex, for Kafka: https://github.com/apache/beam/blob/master/sdks/java/io/kafka/upgrade/src/main/java/org/apache/beam/sdk/io/kafka/upgrade/KafkaIOTranslation.java

Chartboost

© Chartboost. All rights reserved.

What’s Beam YAML & Protobuf

Is a new SDK that uses a declarative approach to creating data processing pipelines using YAML

Protocol Buffers (Protobuf) is a language-neutral, platform-neutral serialization created by Google, enabling efficient and compact data exchange through structured schemas.

Chartboost

© Chartboost. All rights reserved.