BEAM
COLLEGE

Dataflow Templates: Google-Provided Beam Pipelines

Oleg Semenov
Zachary Houfek

Dataflow and Templates

Dataflow

Google Cloud’s native way of running Beam pipelines.

Fully-managed and serverless.

Dataflow Templates

Service for staging Beam pipelines in Google Cloud.

Pipelines always use the Dataflow runner.

There are many Google-supported templates pre-staged for all users.

Why Care?

Why Stage?

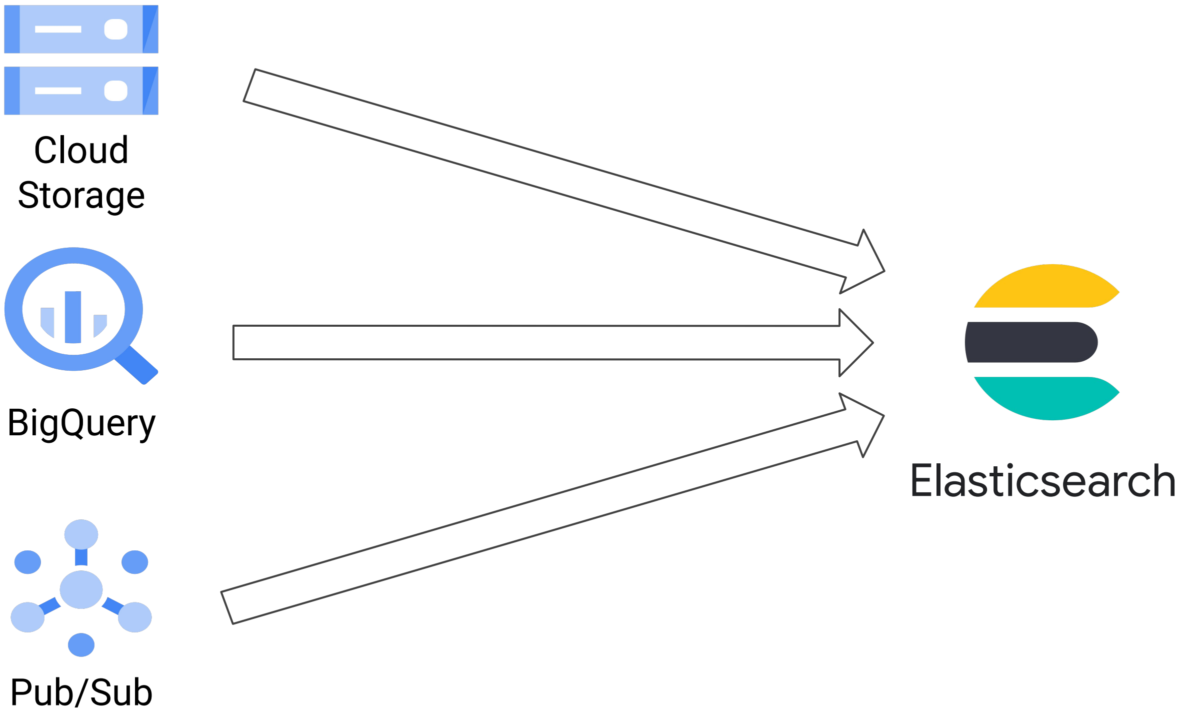
- Easily allow many people to reuse the same pipeline.
- More flexibility in how jobs are launched.
- Currently empowering other Google Cloud products like:
  - Dataplex
  - Datastream
  - Spanner

Why care as a Beam user?

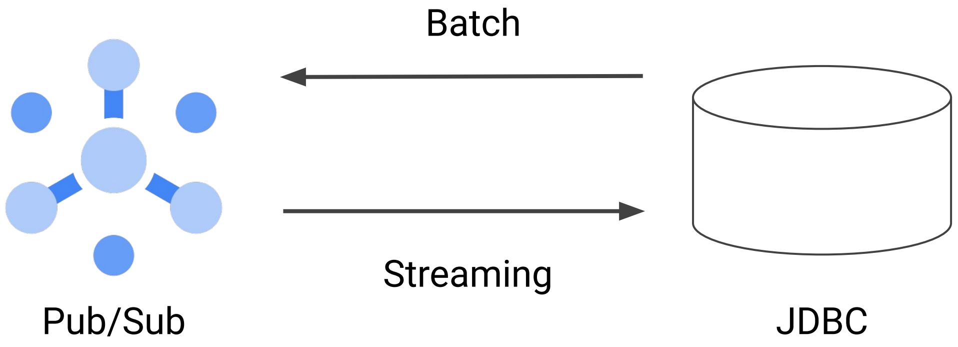
- Dataflow Templates are Beam pipelines.
- Google-provided templates are open source.
  - Many examples of how to write and test pipelines.
  - Serve as a starting point for your own pipelines.
  - github.com/GoogleCloudPlatform/DataflowTemplates

New Templates From 2021-2022

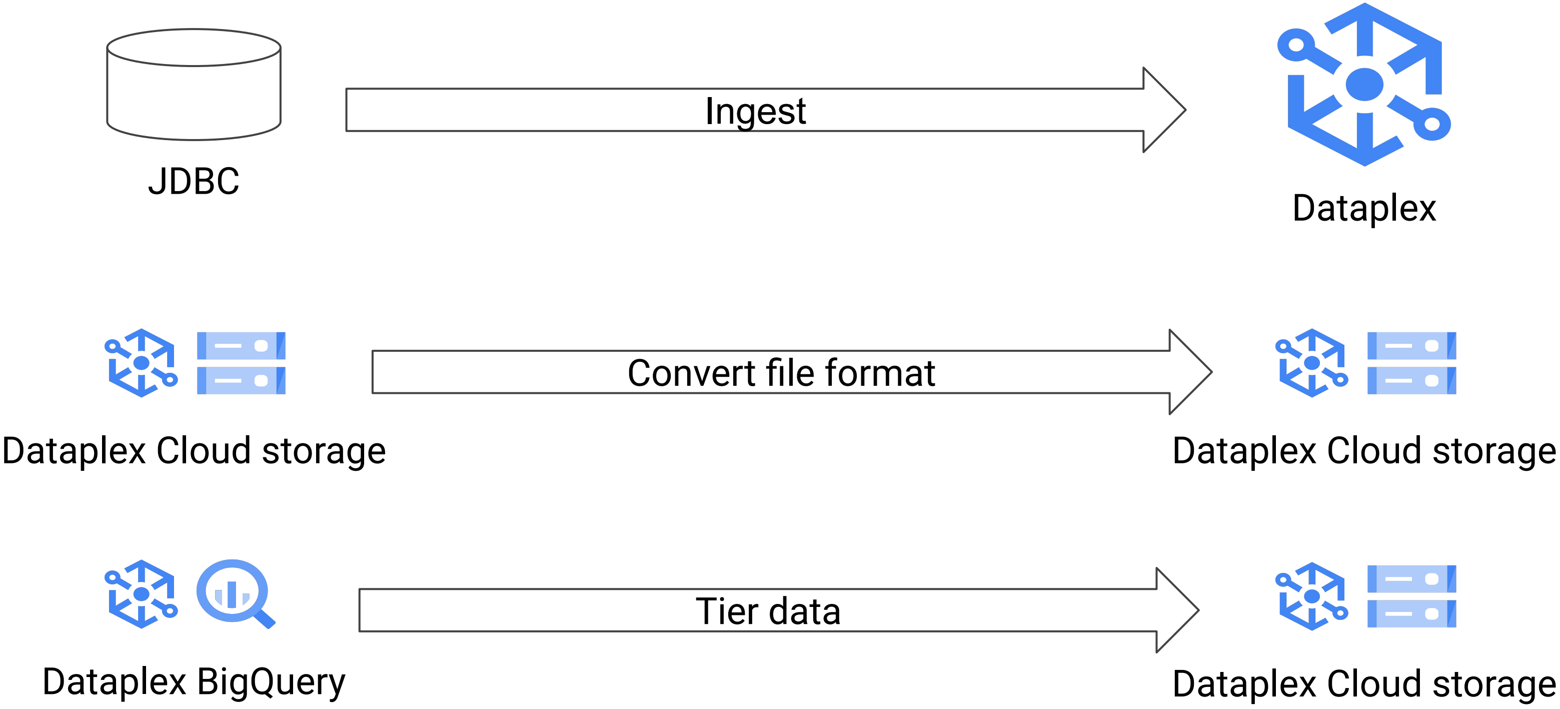
Elasticsearch



Pub/Sub and JDBC



Dataplex



JDBC

Ingest

Dataplex

Dataplex Cloud storage

Convert file format

Dataplex Cloud storage

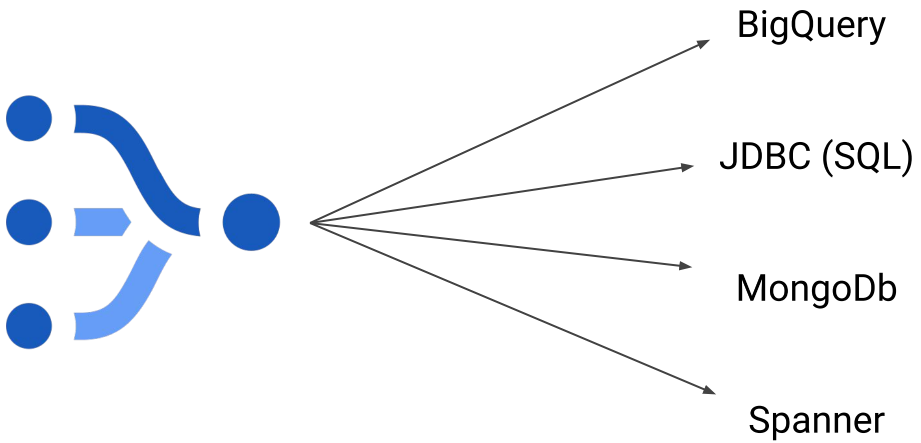
Dataplex BigQuery

Tier data

Dataplex Cloud storage

Demo

Datastream



BigQuery

JDBC (SQL)

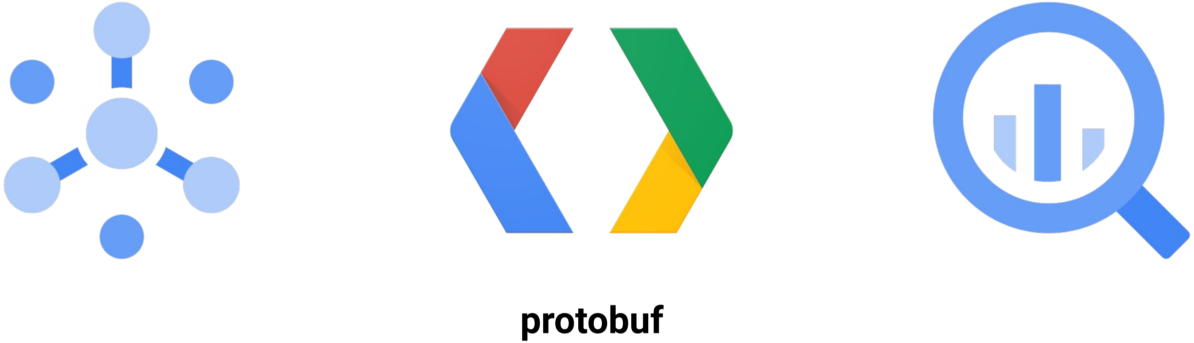
MongoDb

Spanner

Datastream Continued

- DatastreamIO
  - Batch and Streaming from GCS
  - Streaming from Pub/Sub file notifications
- Error handling to GCS (BigQuery and Spanner)
- Stateful processing for filtering and ordering (SQL)
- More available on GitHub

Pub/Sub Proto -> BigQuery



Pub/Sub Proto -> BigQuery Features

Determines type dynamically based on proto descriptor stored in GCS

Modify the input using a JavaScript UDF

Invalid messages sent to dead-letter topic

Can convert the proto schema to a BigQuery schema

Demo

Pub/Sub Proto -> BigQuery Beam Contributions

PubsubIO Proto Support

- Previously required the message type known at compile time.
- Now supports proto DynamicMessage.

PubsubIO Error Handling

- Avro and proto parsing errors could permanently stall pipeline on Dataflow.
- Now can send the error to a dead-letter topic.
- Other improvements:
  - Coder-safe Throwable wrapper
  - PubsubTestClient supports doing pull and publish with same client instance

Unreleased

- BigQuery to Bigtable
- Spanner Change Stream (GCS and BigQuery)

Looking Ahead

Better Engaged on GoogleCloudPlatform/DataflowTemplates

Now in Dataflow Templates

- Googlers tend to develop Dataflow Templates internally and sync externally.
- External contributions to Dataflow Templates go through internal checks.
- Templates team not engaged enough on the Dataflow Templates repository.

Goal for Dataflow Templates

- Templates team working primarily from the Dataflow Templates GitHub repo.
- Templates team responds faster to Issues and PRs.
- Make life easier on external contributors.
  - Better documentation on how to develop and test Dataflow Templates.
  - Open source Templates integration tests.

Thanks to those who contributed through GitHub!

alxavier

ayushpoddar

Billy Jacobson

Diego de Lima

dhercher

johndallard

Kenzyme L

melbrodrigues

Omid Tourzan

Naruhito

Nathan J. Mehl

Panas Cherepushko

pareshsarafmdb

Thank you!

And thanks to all contributors!