

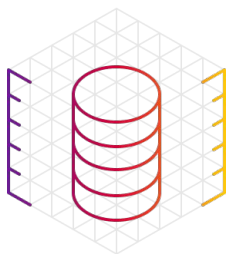
DATASTACK

DEVELOPERS

# Data Analytics with Apache Beam and Google DataFlow

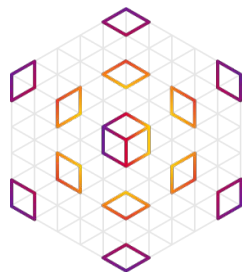
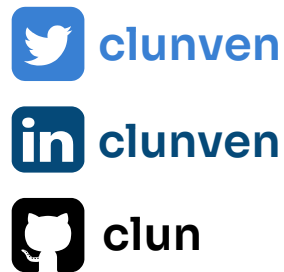
Move your data like a boss





# › Cédric Lunven

## Developer Relations Software Engineering



{ REST }

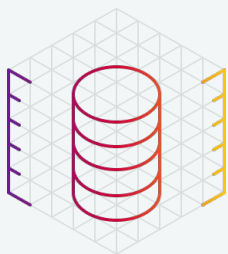
GRPC

- ❖ Trainer
- ❖ Public Speaker
- ❖ Developers Support
- ❖ Developer Applications
- ❖ Developer Tooling
- ❖ Creator of ff4j (ff4j.org)
- ❖ Maintainer for 8 years+
- ❖ Implementing APIs for 8 years

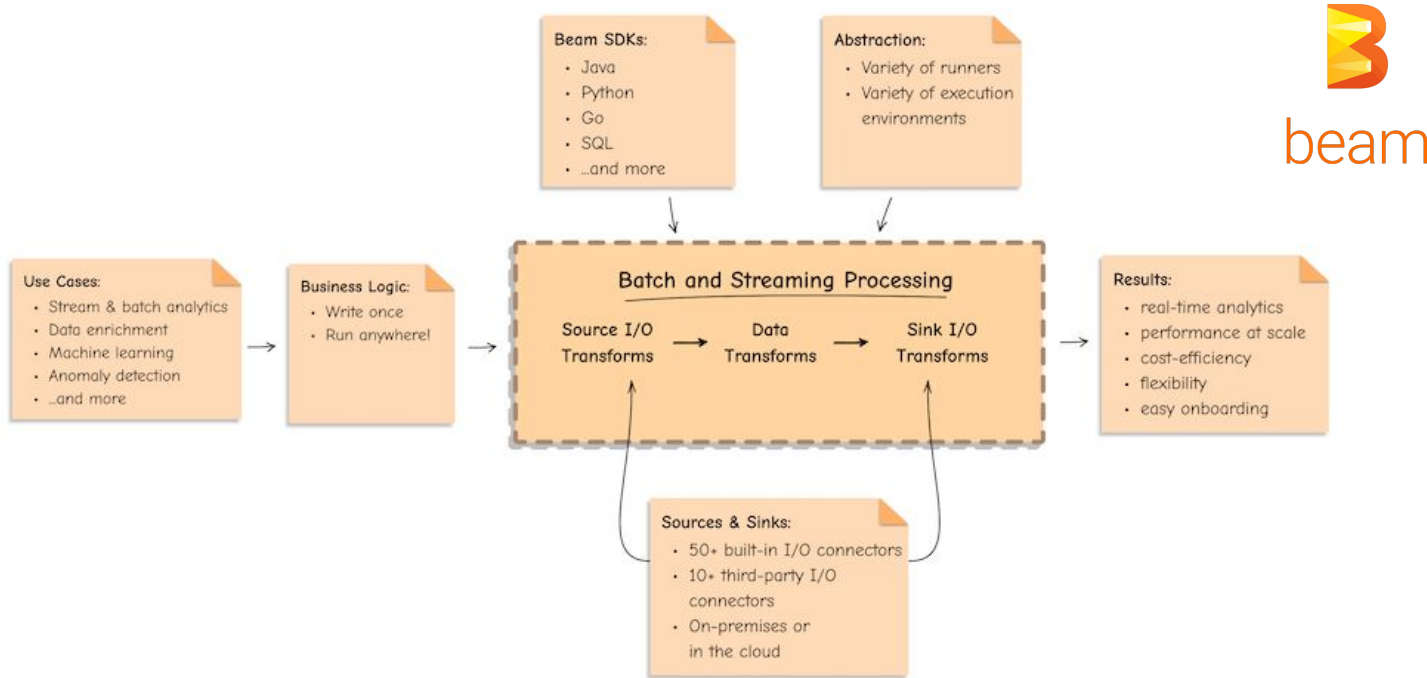
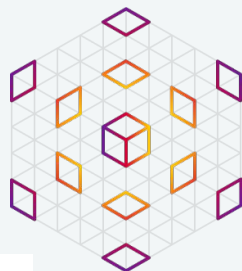


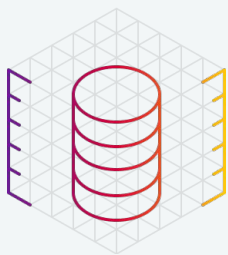
# Agenda (30min)

- › Apache Beam
- › AstraDbIO
- › **LAB Load Csv into Astra**
- › GCP, dataflow runners, templating
- › **Walkthrough more flows**
- › Let's stay in touch !

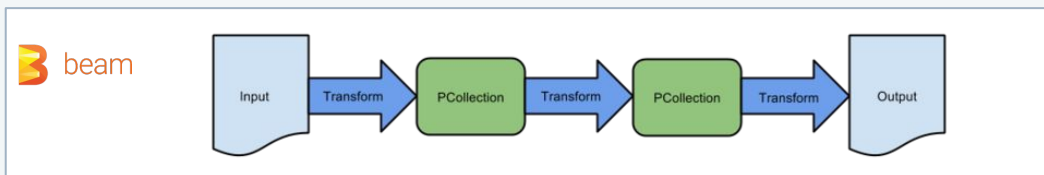
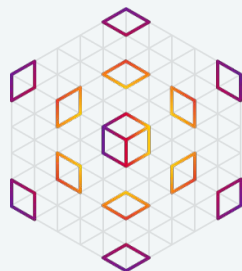


# » Apache Beam

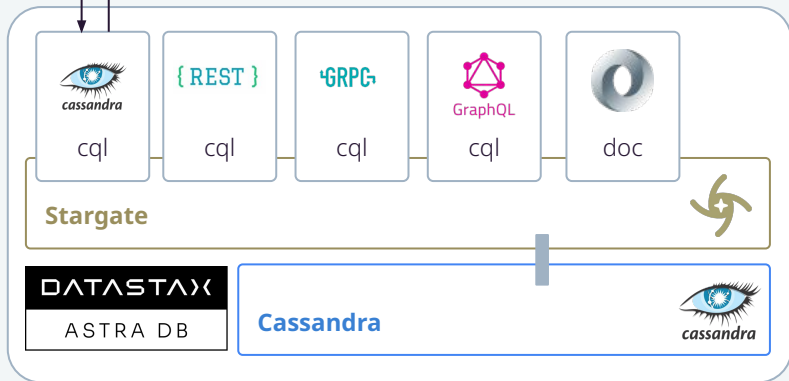




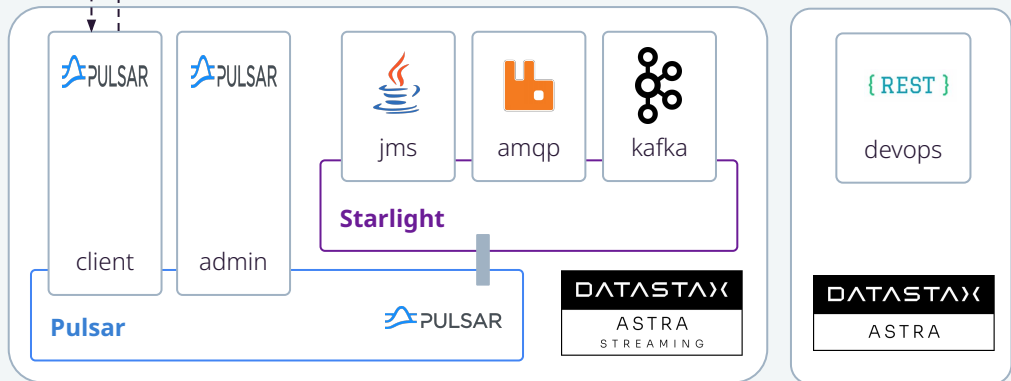
# › Integrating Astra and Beam

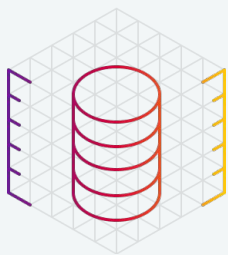


## 1 Bulk Data (*CassandraIO*)

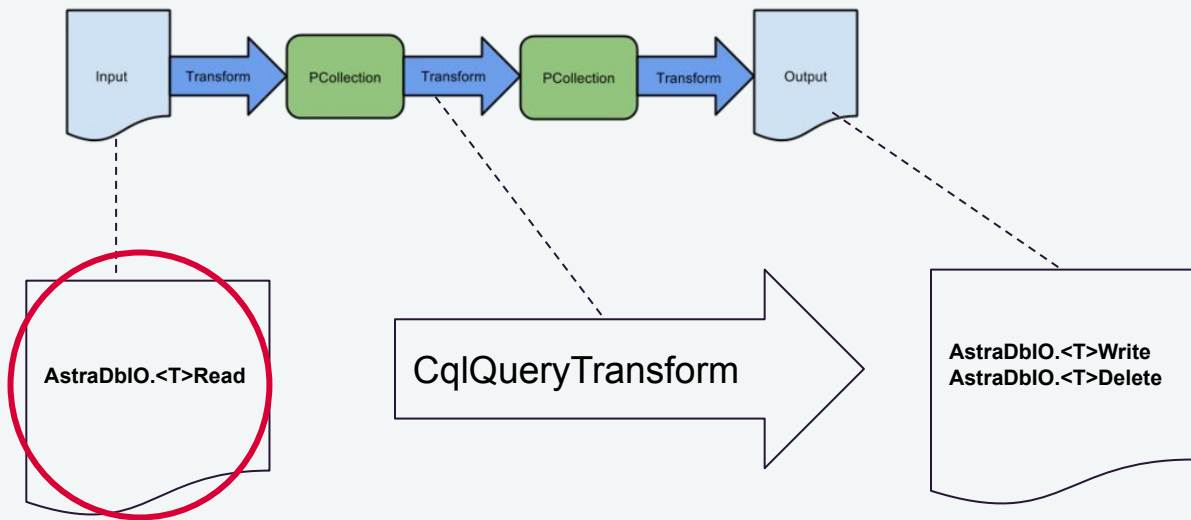
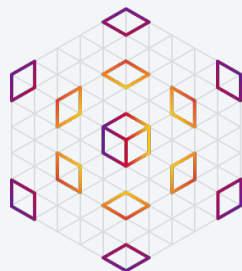


## 2 Streaming Data (*PulsarIO*)





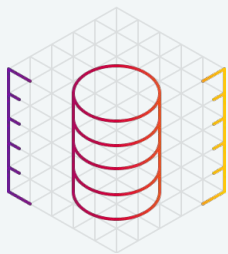
# › Astradblo



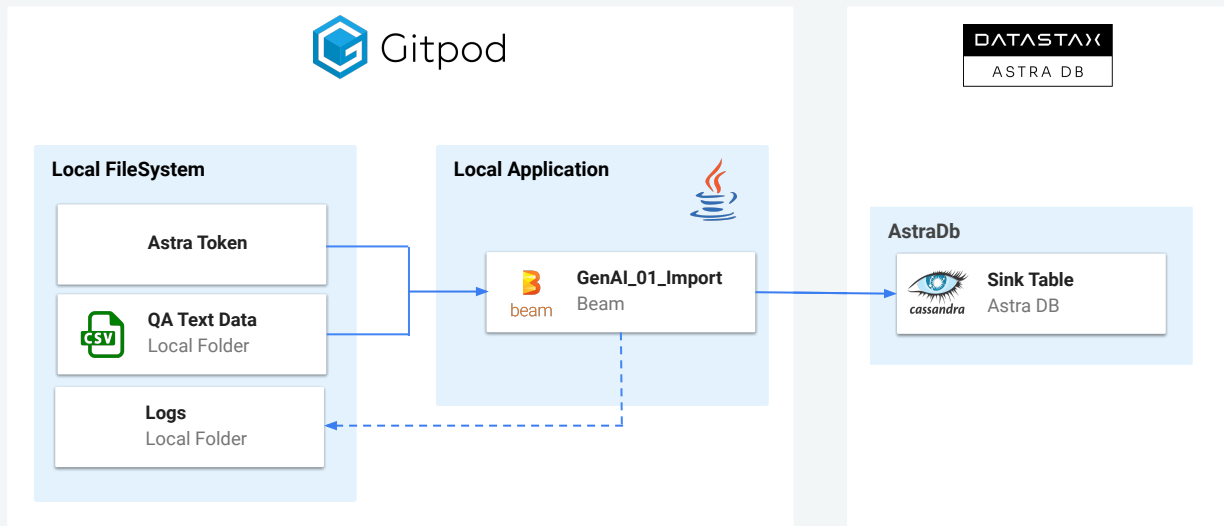
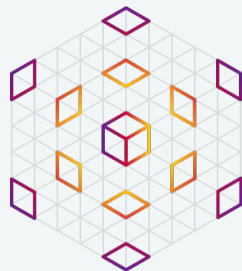
- Token ranged query
- Session Management

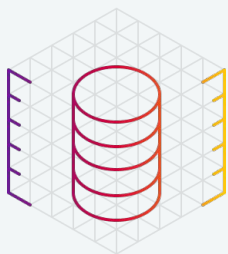
- **Upgrading drivers**
- **New parameters**
- **Enforcing best practices**
- **Enforcing consistency**
- **Dynamic Schema**
- **Dynamic mapping to Row**
- **New Transforms for CQL**
- **Templates for Options**
- **Automatic Table Creation**
- **Enhanced logging**

- Batch operations
- Back Pressure
- Object Mapping

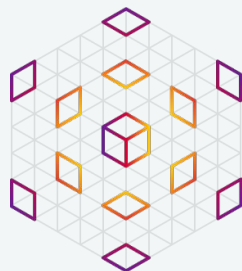



## ➤ Flow #1: Load CSV to into Astra







# Awesome Astra




 **Awesome-Astra**





 awesome-astra/docs  
☆ 26 👤 13



[Home](#) [Get Started](#) [Manage Data](#) [Developers](#) [AI/ML](#) [Tools/Integrations](#) [Sample App Gallery](#) [Contact Us](#)


 **Awesome-Astra**

Tutorials, documentation and learning materials to build great applications or connect external tools to Astra.


 [Get started](#)

[Start Coding](#) 


  


 **DBAAS Apache Cassandra**

Start your cluster in the cloud within a few minutes for free with **no credit card required**. Keep using the free tier as much as you want with 20 million queries a month

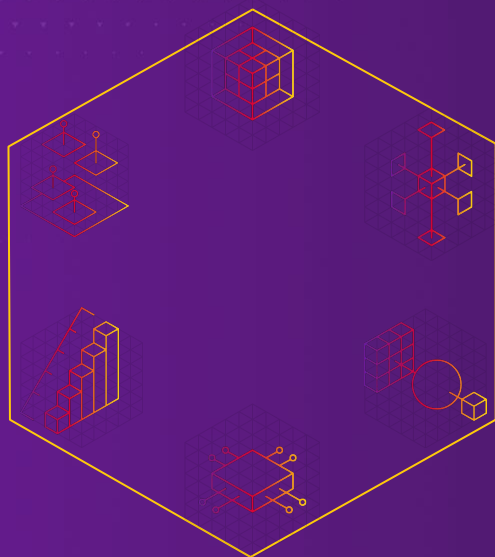
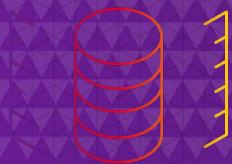
 **Different Apis with Stargate**

Leverage **gRPC**, **graphql**, and **REST** apis to build great stateless applications. You can keep using Cassandra Drivers and related frameworks if you prefer

 **Streaming with Apache Pulsar**

Introduce realtime data in your projects with **Apache Pulsar™** support with lots of interfaces (JMS, AMQP, Kafka) and features provided (functions, CDC, sinks...)





# Lab

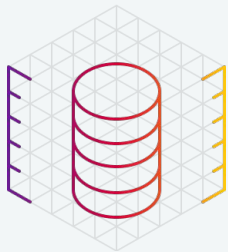
## Load a CSV to Astra

**1.1 Initialize Astra**

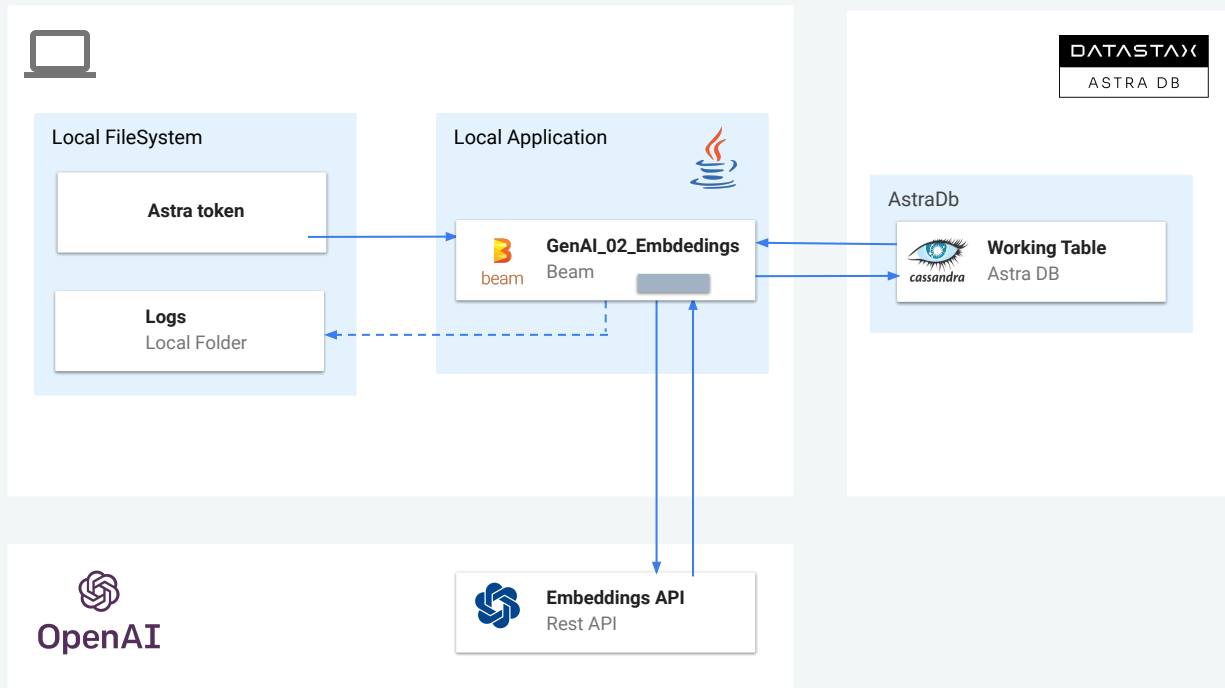
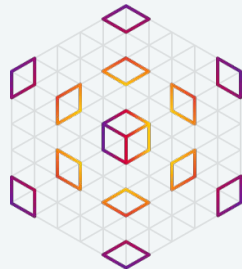
**1.2 Create Astra Token**

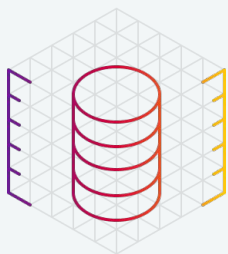
**1.3 Launch Gitpod**

**1.4 Run the flow**

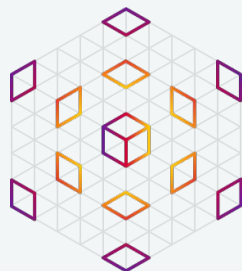


## › Flow #2: Compute Embeddings





# › Apache Beam Runners



Apache Beam  
Direct Runner



Apache Apex



Apache Spark



Apache Flink



Apache Gearpump



Google Cloud  
Dataflow



IBM Streams



Apache Storm  
WIP



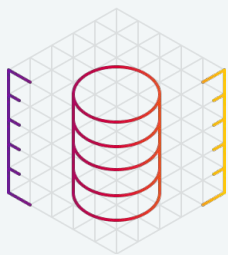
Ali Baba  
JStorm



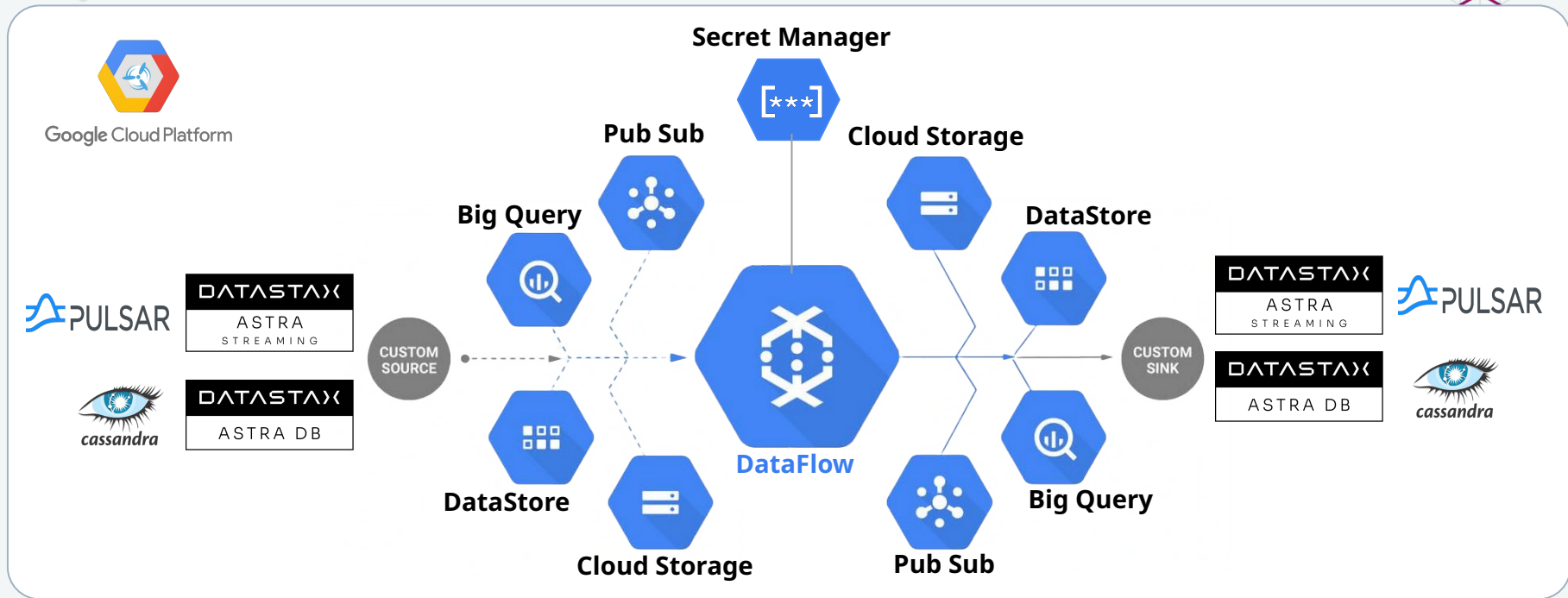
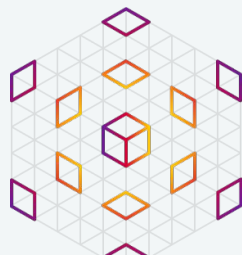
Apache Samza

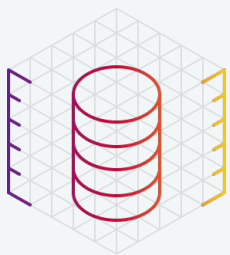


Hadoop  
MapReduce

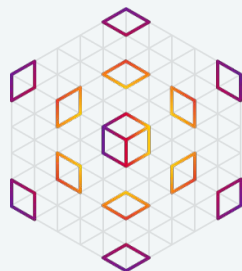


# Google Cloud Platform





# ➤ Flow Astra Db to BigQuery



Google Cloud

Credenti

**Astra Token**  
Secret Manager

DataSet

**Sink Table**  
Big Query

**AstraDb\_To\_BigQuery**  
Dataflow

Operational Data

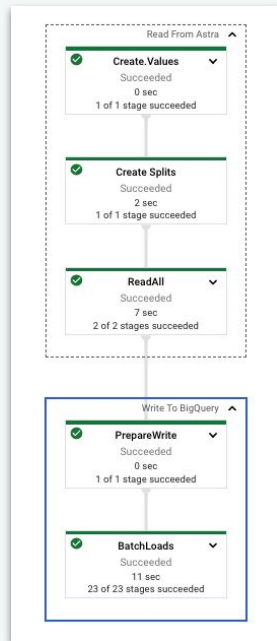
**Temp,Staging,Logs**  
Cloud Storage

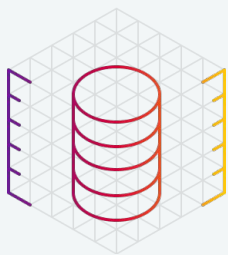
Source

**Source Table**  
Astra DB

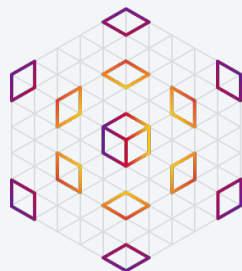
DATASTAX

ASTRA DB





# » Flex Templates



- <https://github.com/GoogleCloudPlatform/DataflowTemplates>
- Dynamic and convention based mapping
  - Cassandra ↔ Beam ↔ Dataflow
- Guided Tutorials
- Troubleshooting guide
- AstraDb to bigquery
- BigQuery to AstraDb
  - Which table would you create ?

**Google Cloud** Integrations Search (/) for resources, docs, products, and more

**Dataflow** Create job from template

Job name \*  
My first job  
Must be unique among running jobs

Regional endpoint \*  
us-central1 ( Iowa )  
Choose a Dataflow regional endpoint to deploy worker instances and store job metadata. You can optionally deploy worker instances to any available Google Cloud region or zone by using the worker region or worker zone parameters. Job metadata is always stored in the Dataflow regional endpoint. [Learn more](#)

Default template \*  
Cassandra to Cloud Bigtable  
A pipeline to import a Apache Cassandra table into Cloud Bigtable. [OPEN TUTORIAL](#)

**Required Parameters**

Cassandra Hosts \*  
Comma separated value list of hostnames or ips of the Cassandra nodes.

Cassandra Keyspace \*  
Cassandra Keyspace where the table to be migrated can be located.

Cassandra Table \*  
The name of the Cassandra table to Migrate

Bigtable Project ID \*  
The Project ID where the target Bigtable Instance is running.

Target Bigtable Instance \*  
The target Bigtable Instance where you want to write the data.

Target Bigtable Table \*  
The target Bigtable table where you want to write the data.

Temporary location \* **BROWSE**  
Path and filename prefix for writing temporary files. Ex: gs://your-bucket/temp

**Encryption**

☒ Google-managed encryption key  
No configuration required

☐ Customer-managed encryption key (CMK)  
Manage via [Google Cloud Key Management Service](#)

**Additional information**

Read from Cassandra

Convert Row

Write to Bigtable

Release Notes

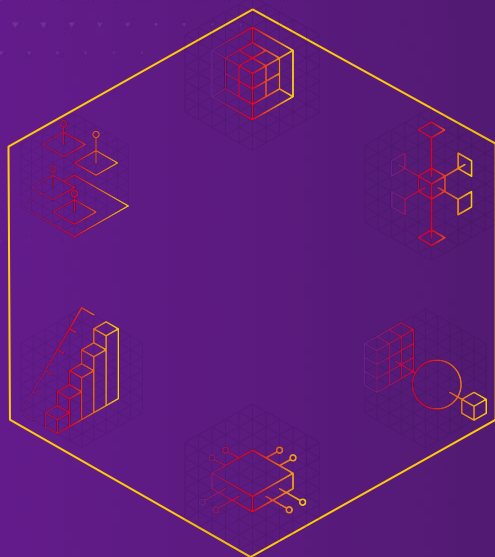


# WalkThrough

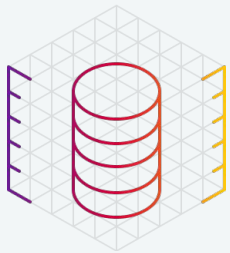
**2.1 Generative AI**

**2.2 Copy data to GCP**

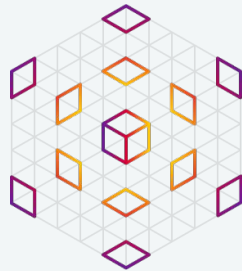
**2.3 Flex Template**



[github.com/datastaxdevs/workshop-beam](https://github.com/datastaxdevs/workshop-beam)



## ➤ What's Next ?



**Discord:** [dtsx.io/discord](https://dtsx.io/discord)

**Academy:** [academy.datastax.com](https://academy.datastax.com)

**Workshops:** [datastax.com/workshops](https://datastax.com/workshops)

**YouTube:** @DataStax Developers



DATASIX



# Thank You