



Developing Beam IOs: CDAP and SparkReceiver

By Alex Kosolapov and Elizaveta Lomteva

Agenda

- Introduction
- Developing an IO
- CDAP IO Overview
- Streaming Source IO – SparkReceiver
- Testing IO
- Akvelon Data Analytics and ML Accelerators demo

AKV₃ELON

1200+
technology experts

23+
years of expertise

150+
clients

15 offices
in 11 countries

24/7
operations support





AKVELON



Google Cloud
Partner

Developing Beam IO (Java)

- Starting point: [Developing a new I/O connector](#)
- Design:
 - Define the input/output format
 - Read – Splittable DoFn (SDF), Write – ParDo
 - Determine target pipeline configuration parameters
- Develop:
 - DoFn to process an element
 - Read/Write PTransforms
- Test IO:
 - Unit testing, Integration, Performance testing
- Release: IO Documentation and examples



An open-source platform for data applications in
hybrid and multi-cloud environments

Visual point-and-click interface enabling code-free
deployment of ETL/ELT data pipelines

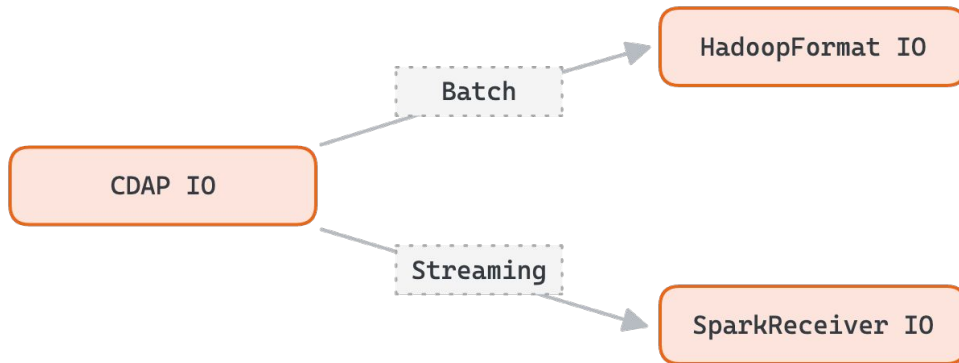
Ecosystem of plugins, including business
applications connectors

CDAP IO

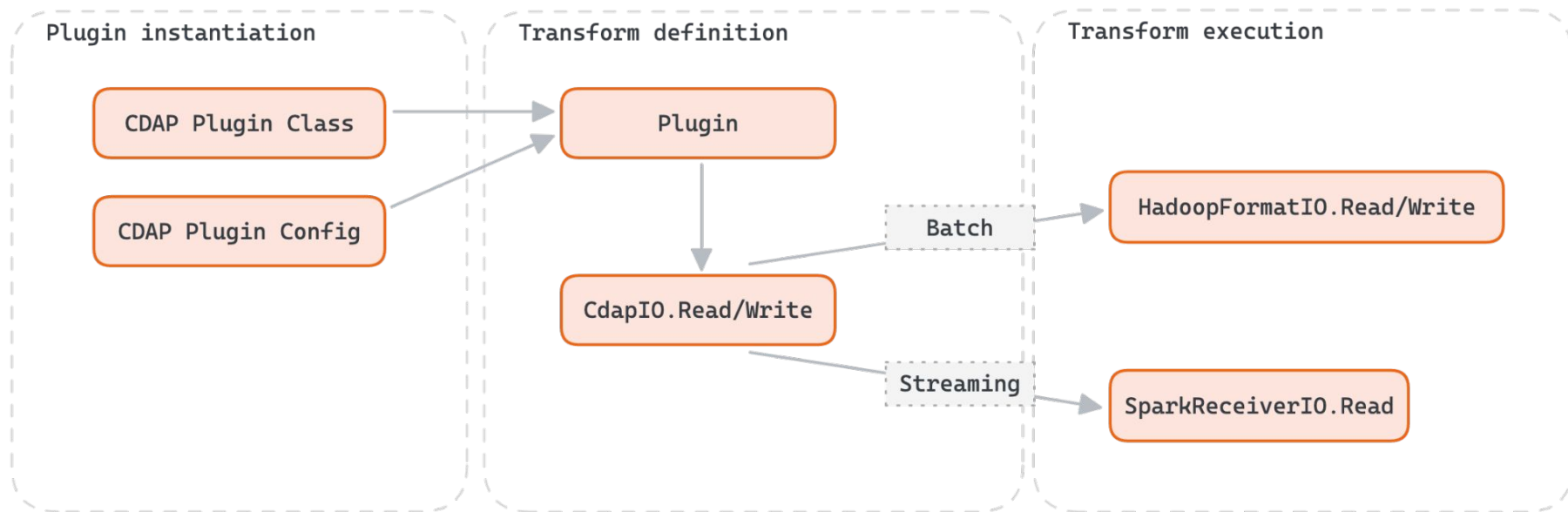
Provides transforms for reading and writing data via CDAP plugins

Connects Apache Beam with a variety of business applications like Salesforce, Hubspot, ServiceNow and Zendesk

Uses CDAP plugin definition



CDAP IO Workflow



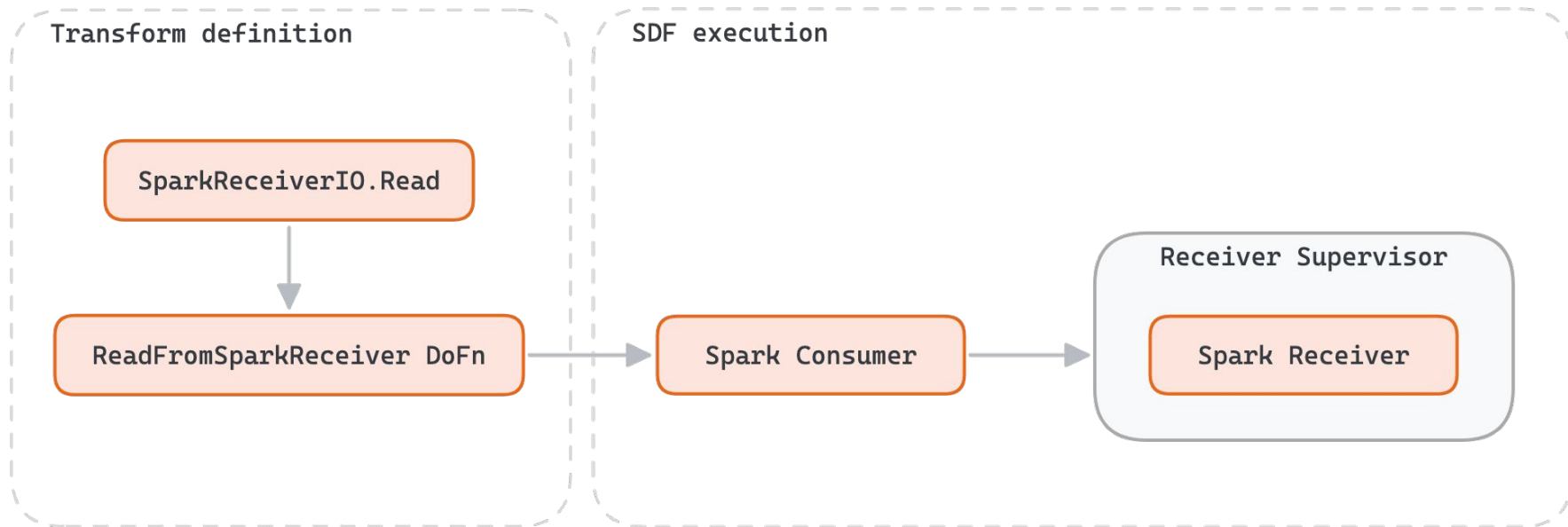
SparkReceiver IO

SparkReceiverIO provides transforms to read data via Apache Spark Receiver

Prerequisites:

- Spark Receiver provides HasOffset interface.
- Records have a numeric field that represents record offset.

SparkReceiver IO Workflow



Beam Parallelism & IO

Input parallelism – reading from bounded and unbounded sources, i.e. data source parallelism

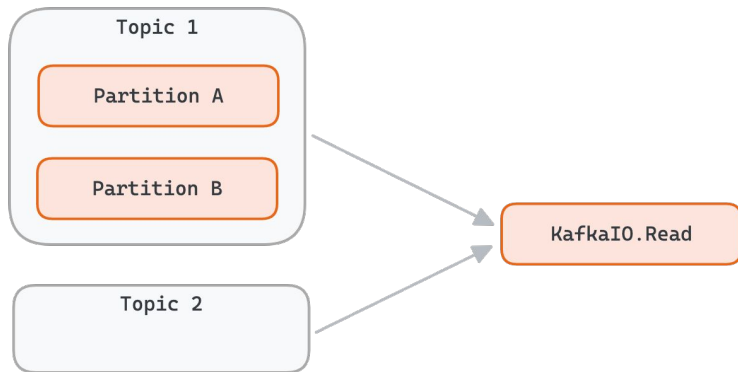
Inter-stage parallelism – splitting processing across workers, e.g. key-based data partitioning

Intra-stage parallelism – splitting element processing within transforms, e.g. Splittable DoFns, bundle processing

Data Source Parallelism

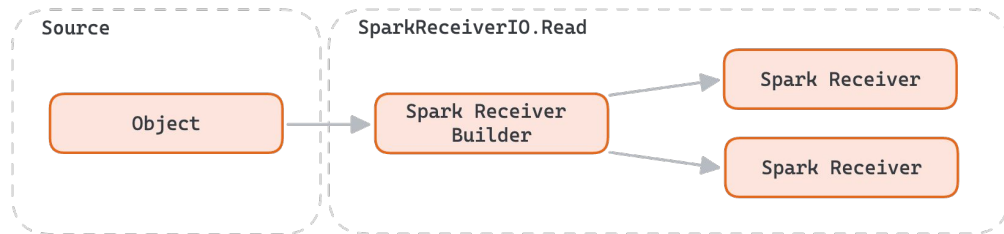
Refers to the parallelism achieved by reading data from multiple sources or partitions of a single source concurrently.

(E.g. Kafka topic partitions)



SparkReceiverIO

Each receiver builder can be associated with single source object and create multiple receivers during processing



Inter-stage parallelism

Refers to the parallelism between different transforms (or stages) within a Beam pipeline.

Achieved by runner implementation

(E.g. key-based operations in Beam)

SparkReceiverIO

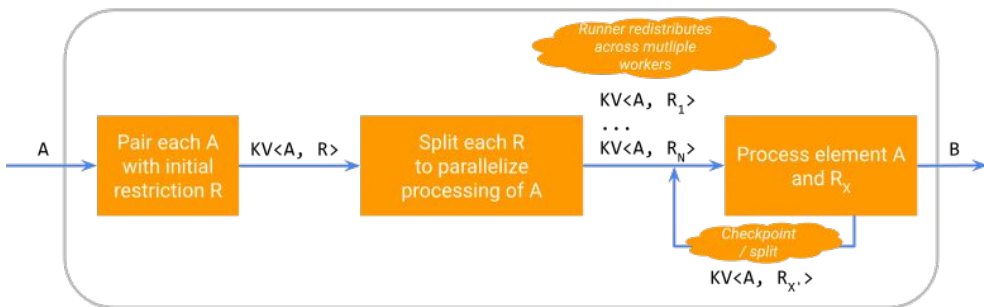
Achieved by supported runners – Direct runner and Dataflow runner v1 and v2



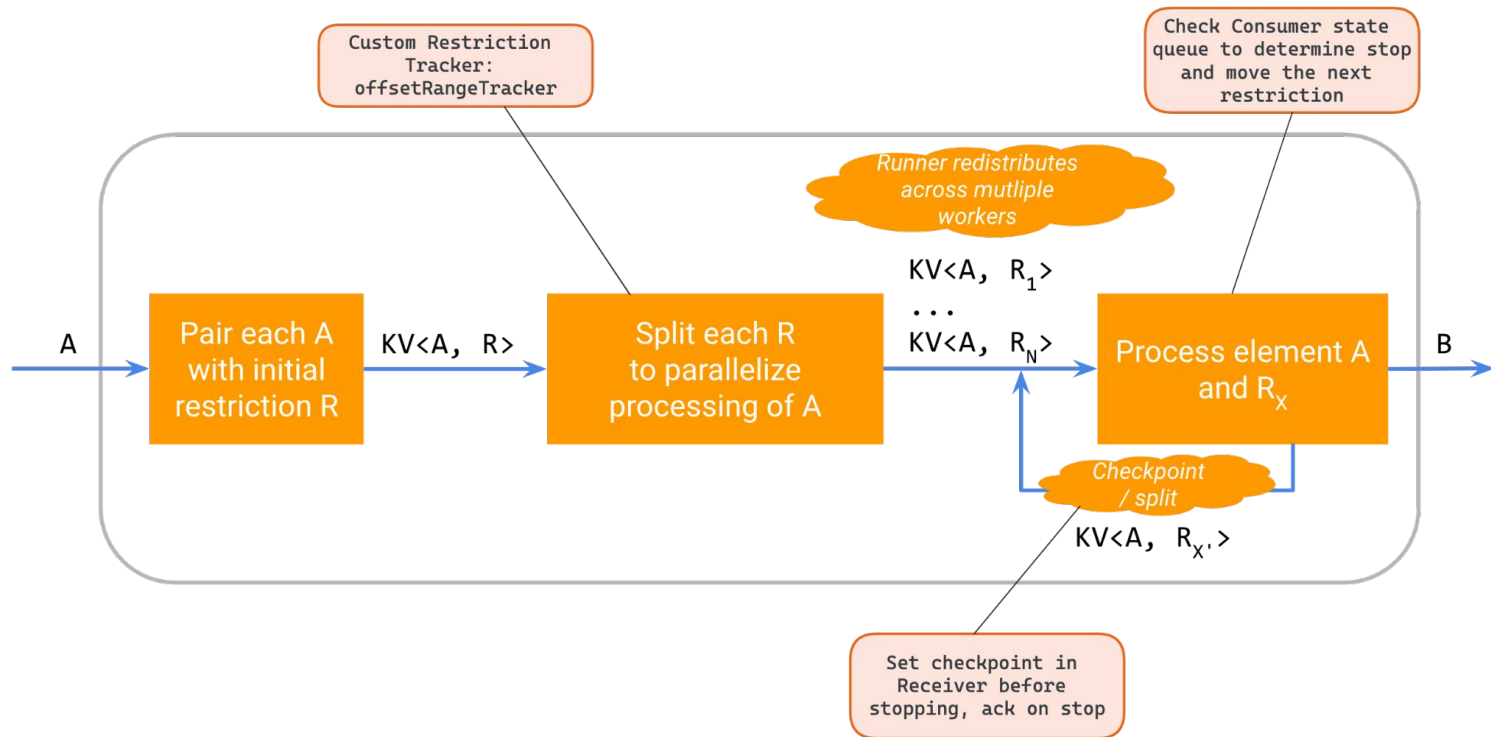
Intra-stage: Splittable DoFn (SDF)

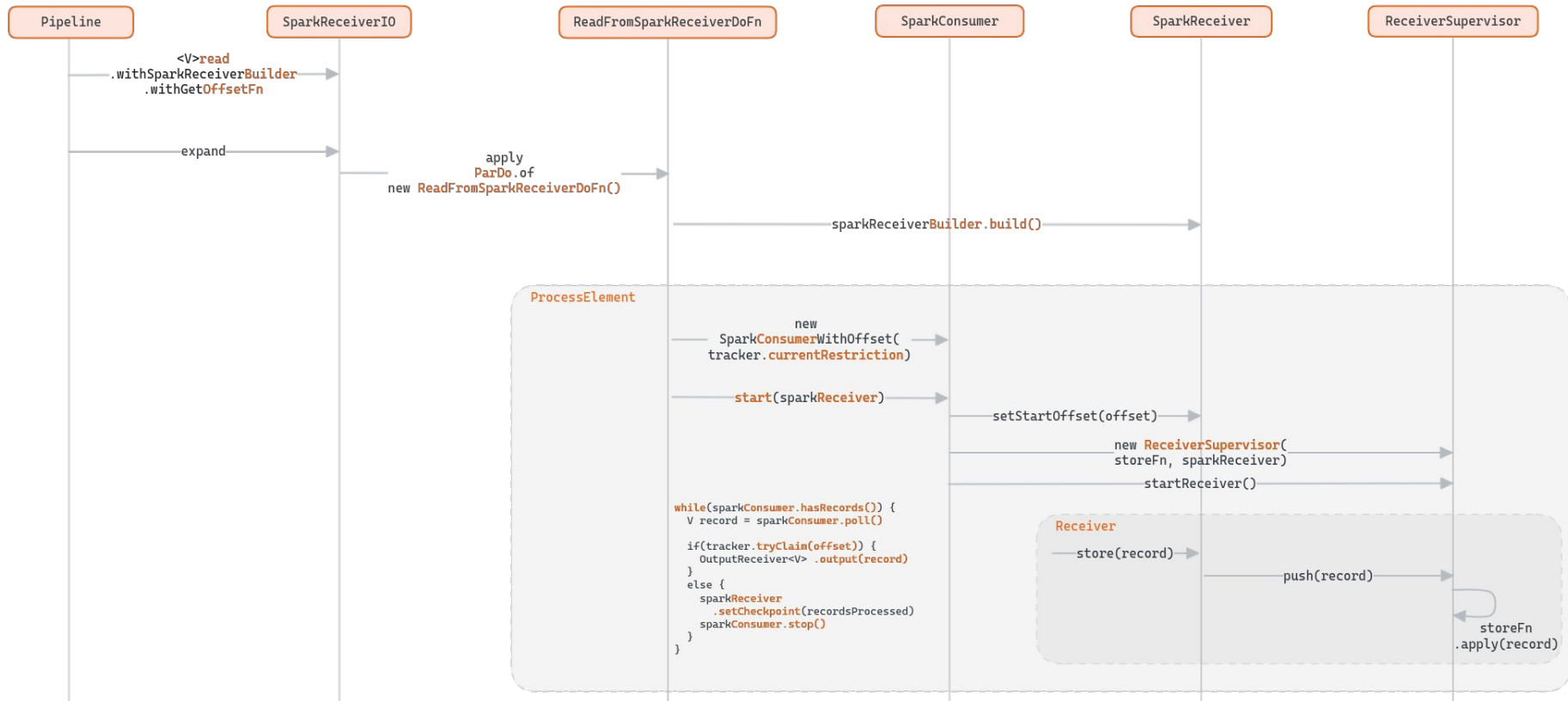
Executing an SDF follows the following steps:

1. Each **element** is paired with a **restriction** (e.g. filename is paired with offset range representing the whole file).
2. Each element and restriction pair is **split** (e.g. **offset** ranges are broken up into smaller pieces).
3. The runner redistributes the element and restriction pairs to several workers.
4. Element and restriction pairs are processed in parallel (e.g. the file is read). Within this last step, the element and restriction pair can pause its own processing and/or be split into further element and restriction pairs.



SparkReceiverIO





Agenda

- Introduction
- Developing an IO
- CDAP IO Overview
- Streaming Source IO – SparkReceiver
- **Testing IO**
- Akvelon Data Analytics and ML Accelerators demo

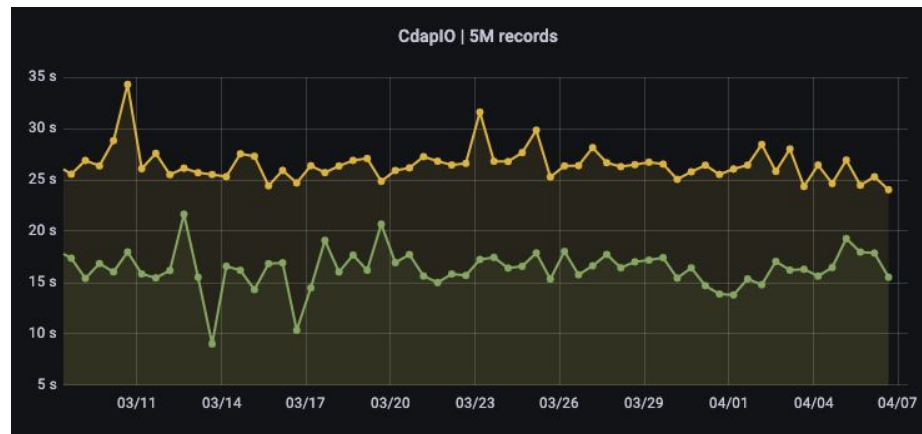
Testing IO and Release

IO Testing

- [testing guide](#), [IO transforms testing](#)
- Unit, integration and [performance test](#)
- Created RabbitMQ SparkReceiver on-demand source in Apache Beam that generates streaming data according to provided profile

Release

- Beam website [IO Connectors](#)
- Documentation & Readmes
- [Complete examples](#)



Documentation

- Using the Documentation
- Concepts
 - Beam programming guide
- Pipelines
- PCollections
- Transforms
- Pipeline I/O
 - Using I/O transforms
 - I/O connectors
 - I/O connector guide
 - Apache Parquet I/O connector
 - Hadoop Input/Output Format I/O
 - HCatalog I/O
 - Single Pigberry I/O connector
 - Scalable I/O connector
 - CDAP I/O connector
 - Spark Receiver I/O connector
 - SingleRowIO I/O connector
 - Developing new I/O connectors
 - Testing I/O transforms

Cdap IO

A `CdapIO` is a transform for reading data from source or writing data to sink CDAP plugin.

Batch plugins support

`CdapIO` currently supports the following CDAP Batch plugins by referencing `CDAP_PLUGIN` class name:

- [Hubspot Batch Source](#)
- [Hubspot Batch Sink](#)
- [Salesforce Batch Source](#)
- [Salesforce Batch Sink](#)
- [ServiceNow Batch Source](#)
- [Zendesk Batch Source](#)

Also, any other CDAP Batch plugin based on Hadoop's `InputFormat` or `OutputFormat` can be used. They can be easily added to the list of supported by class name plugins, for more details please see [CdapIO README](#).

Streaming plugins support

`CdapIO` currently supports CDAP Streaming plugins based on [Apache Spark Receiver](#).

Requirements for CDAP Streaming plugins:

Apache Beam pipeline examples for CdapIO and CDAP plugins

This directory contains set of [Apache Beam](#) pipeline examples to read data from a [CDAP plugin](#) and write data into .txt file (and)

Supported CDAP plugins:

- [ServiceNow](#). More info in the [ServiceNow example README](#).
- [Salesforce](#). More info in the [Salesforce example README](#).
- [Hubspot](#). More info in the [Hubspot example README](#).
- [Zendesk](#). More info in the [Zendesk example README](#).

Demo

AKVELON

Data and Analytics Accelerators

https://github.com/akvelon/DnA_accelerators

Akvelon Data and Analytics Accelerators

Akvelon is a digital product and software engineering company that empowers strategic advantage and accelerates your path to value in Data and Analytics, AI/ML, MLOps, Application development, and more with innovation and predictable delivery. Akvelon is providing this collection of accelerators as a reference and easy customizations for developers looking to build data, machine learning, and visualizations.

- [Get in touch about Data and Analytics and Data Migrations projects.](#)
- [Get in touch about ML projects.](#)
- [Get in touch about Google Cloud projects.](#)

Learn more about all our ML and software engineering services at our website akvelon.com.

Accelerators

ML, Streaming and Batch Data Processing

Apache Beam and Google Cloud Dataflow

[Apache Beam](#) provide unified streaming and batch processing to power ML and streaming analytics use cases. [Google Cloud Dataflow](#) is a managed to run Apache Beam in cloud with minimal latency and costs, and integrations with other Google Cloud products like [Vertex AI](#) and [Tensorflow TFX](#). Akvelon, a [Google Cloud Service Partner](#), and an active Apache Beam contributor and [Beam Summit](#) partner, presents several of our favorite accelerators for Dataflow.

- [Salesforce to Txt](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow, using Apache Beam [CDAP IO](#).
- [Salesforce to BigQuery](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow and BigQuery, using Apache Beam [CDAP IO](#). Flex templates provide a comprehensive example of using Machine Learning (ML) to process streaming data in Dataflow, using Java multilanguage pipeline with Python transforms to run custom TFX and PyTorch [ML models](#). This complete Flex template example also demonstrates creating and setting up [Expansion Service](#) in Dataflow to enable running custom Python transforms within a Java pipeline.
- [Tensorflow TFX model training with Apache Beam](#) - a Python notebook and Python Beam pipeline that demonstrates both Jupyter notebook to train a Tensorflow TFX ML model and the converted Python pipeline ready for Expansion Service use
- [PyTorch ML model training and Expansion Service for multilanguage pipelines with Apache Beam](#) - a complete example to train a PyTorch ML model using Apache Beam, convert the notebook to the Python pipeline, create custom Python Transforms and deploy as Apache Beam Expansion Service for Google Cloud Dataflow.

Custom Visualizations

Akvelon has accumulated vast experience with data analytics, custom visualizations, dashboards, and reports for a wide range of industries and use cases. Here are some of our favorite visualization accelerators.

Looker Visuals

github.com/akvelon/DnA_accelerators/tree/main/dataflow

main DnA_accelerators / dataflow

Google Cloud Dataflow Accelerators

Apache Beam provides unified streaming and batch processing to power ML and streaming analytics use cases. Google Cloud Dataflow is managed to run Apache Beam in the cloud with minimal latency and costs, and integrates with other Google Cloud products like Vertex AI and Tensorflow TFX. Akvelon, a Google Cloud Service Partner, and an active Apache Beam contributor and Beam Summit partner, presents several of our favorite accelerators for Dataflow.

Akvelon, a Google Cloud Partner, is providing this open-source collection of Dataflow Flex templates as a reference and easy customizations for developers looking to build streaming, batch, multilanguage data pipelines with ML processing in Google Cloud Dataflow.


Flex Templates for Google Cloud Dataflow

Google Cloud Dataflow Flex Templates are a powerful way to build and run data pipelines on Google Cloud Platform. With Flex Templates, you can package your pipeline code and dependencies as a Docker image, and then run it on Dataflow with just a few clicks. This makes it easy to build and deploy complex pipelines quickly and reliably.

- [Salesforce to Txt](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow, using CDAP IO.
- [Salesforce to BigQuery](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow using Apache Beam CDAP IO. Flex templates provide a comprehensive example of using Machine Learning (ML) to process data in Dataflow, using Java multilanguage pipeline with Python transforms to run custom TFX and PyTorch ML models. Flex template example also demonstrates creating and setting up [Expansion Service](#) in Dataflow to enable running custom transforms within a Java pipeline.

Machine Learning with Google Cloud Dataflow

- [Tensorflow TFX model training with Apache Beam](#) - a Python notebook and Python Beam pipeline that demonstrate both



Summary

Developing Beam IOs

Machine Learning

Multilanguage pipelines

https://github.com/akvelon/DnA_accelerators



https://github.com/akvelon/DnA_accelerators/

README.md

Akvelon Data and Analytics Accelerators

Akvelon is a digital product and software engineering company that empowers strategic advantage and accelerates your path to value in Data and Analytics, AI/ML, MLOps, Application development, and more with innovation and predictable delivery. Akvelon is providing this collection of accelerators as a reference and easy customizations for developers looking to build data, machine learning, and visualizations.

- [Get in touch about Data and Analytics and Data Migrations projects.](#)
- [Get in touch about ML projects.](#)
- [Get in touch about Google Cloud projects.](#)

Learn more about all our ML and software engineering services at our website akvelon.com.

Accelerators

ML, Streaming and Batch Data Processing

Apache Beam and Google Cloud Dataflow

Apache Beam provide unified streaming and batch processing to power ML and streaming analytics use cases. Google Cloud Dataflow is a managed to run Apache Beam in cloud with minimal latency and costs, and integrations with other Google Cloud products like Vertex AI and TensorFlow TFX. Akvelon, a Google Cloud Service Partner, and an active Apache Beam contributor and Beam Summit partner, presents several of our favorite accelerators for Dataflow.

- [Salesforce to Txt](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow, using Apache Beam CDAP IO.
- [Salesforce to BigQuery](#) - Flex templates for batch and streaming Salesforce data processing with Google Cloud Dataflow and BigQuery, using Apache Beam CDAP IO. Flex templates provide a comprehensive example of using Machine Learning (ML) to process streaming data in Dataflow, using Java multilanguage pipeline with Python transforms to run custom TFX and PyTorch ML models. This complete Flex template example also demonstrates creating and setting up Expansion Service in Dataflow to enable running custom Python transforms within a Java pipeline.
- [Tensorflow TFX model training with Apache Beam](#) - a Python notebook and Python Beam pipeline that demonstrates both Jupyter notebook to train a Tensorflow TFX ML model and the converted Python pipeline ready for Expansion Service use
- [PyTorch ML model training and Expansion Service for multilanguage pipelines with Apache Beam](#) - a complete example to train a PyTorch ML model using Apache Beam, convert the notebook to the Python pipeline, create custom Python Transforms and deploy as Apache Beam Expansion Service for Google Cloud Dataflow.

Custom Visualizations

Akvelon has accumulated vast experience with data analytics, custom visualizations, dashboards, and reports for a wide range of industries and use cases. Here are some of our favorite visualization accelerators.

Looker Visuals

AKVELON

https://github.com/akvelon/DnA_accelerators

<https://akvelon.com>

Questions?

<https://www.linkedin.com/in/akosolapov>
<https://www.linkedin.com/in/elizaveta-lomteva>

BEAM
SUMMIT

