Multi-language pipelines

David Sabater Dinter
Product Manager, Google Cloud

Multi-language pipelines

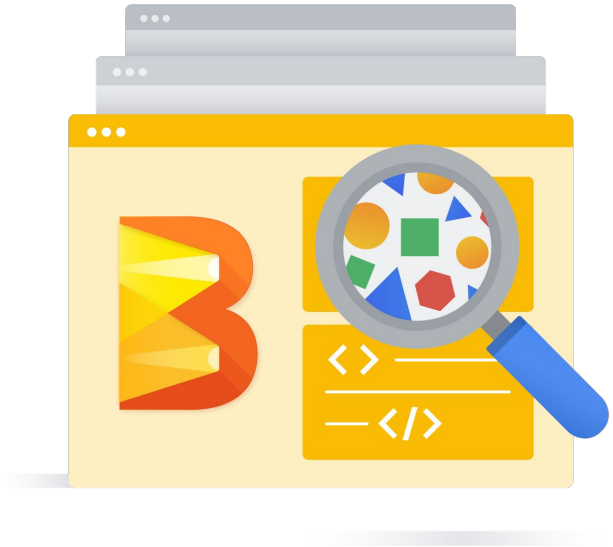
Agenda

Beam portability

Runner v2

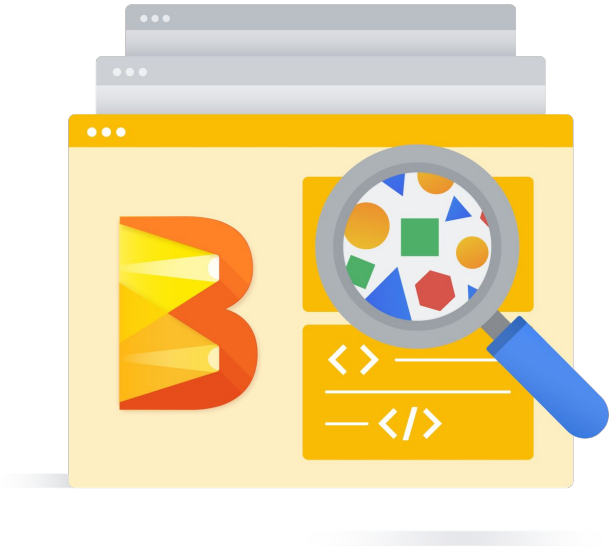
Cross-language transforms

The Beam vision

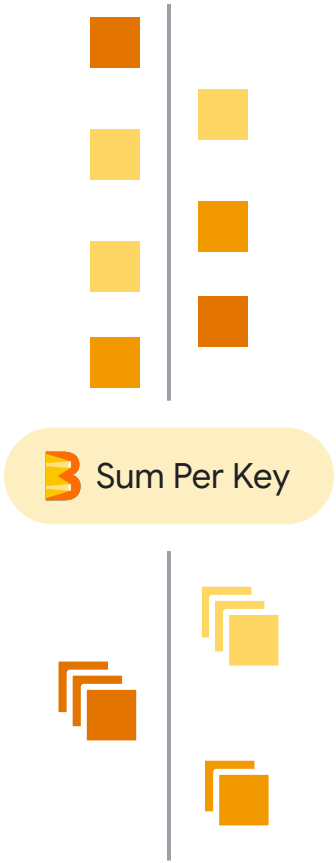


The Beam vision

Provide a **comprehensive portability framework** for data processing pipelines; one that allows you to write your pipeline once in the **programming language of choice** and run it, with minimal effort, on the **execution engine of choice**.



The Beam vision



The Beam vision

Java

Input.apply
(Sum.integersPerKey())

Python

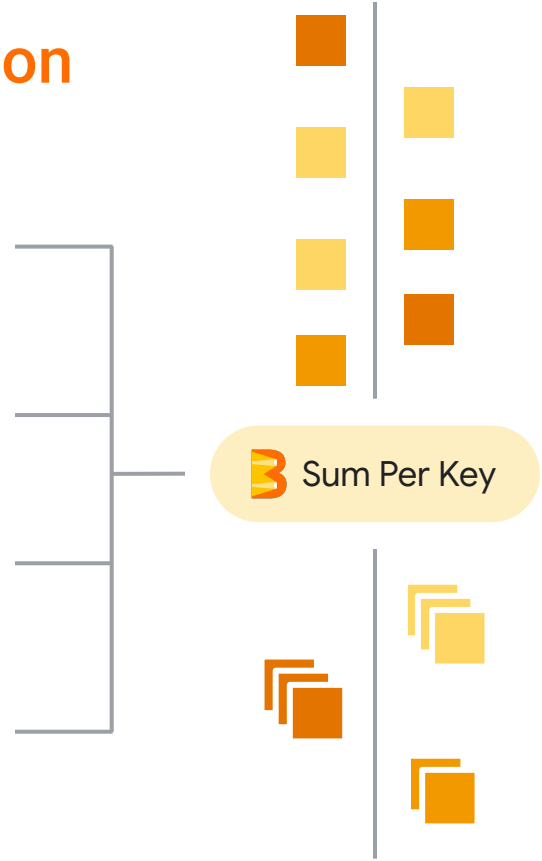
input | Sum.PerKey()

Go

stats.Sum(s, input)

SQL

SELECT key, SUM(value)
FROM input GROUP BY key



The Beam vision

Java

```
Input.apply
(Sum.integersPerKey())
```

Python

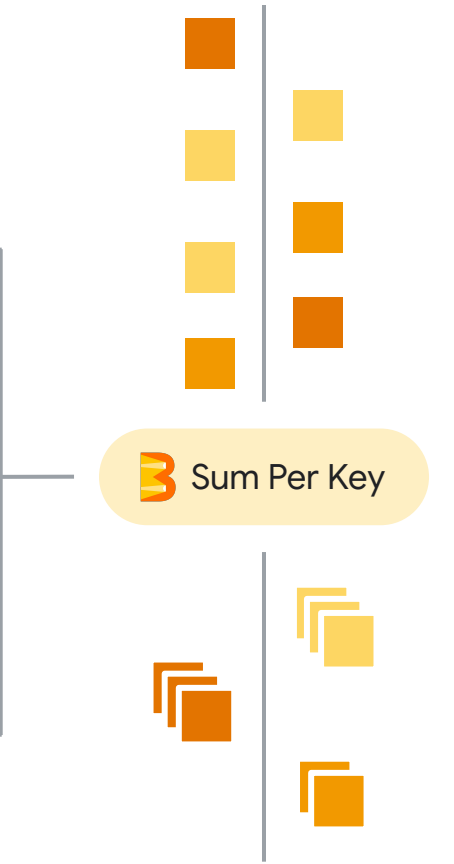
```
input | Sum.PerKey()
```

Go

```
stats.Sum(s, input)
```

SQL

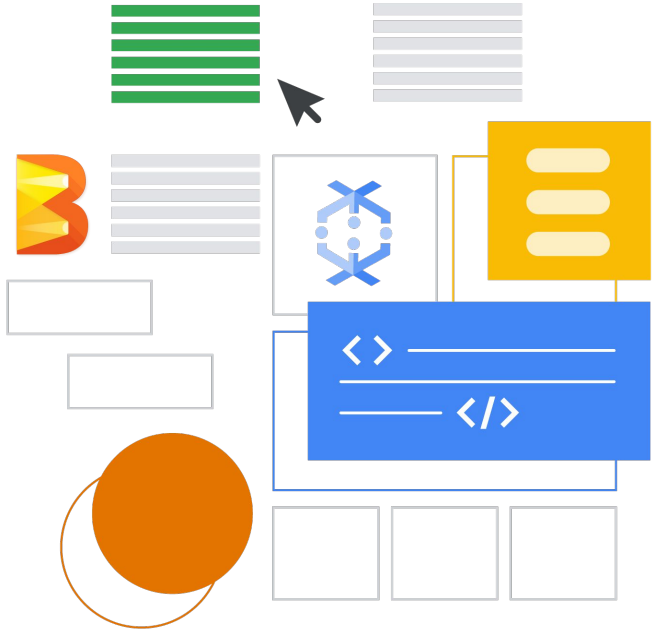
```
SELECT key, SUM(value)
FROM input GROUP BY key
```



- Dataflow
- Apache Flink
- Apache Spark
- Apache Apex
- Gearpump
- IBM Streams
- Apache Samza
- Apache Nemo (Incubating)

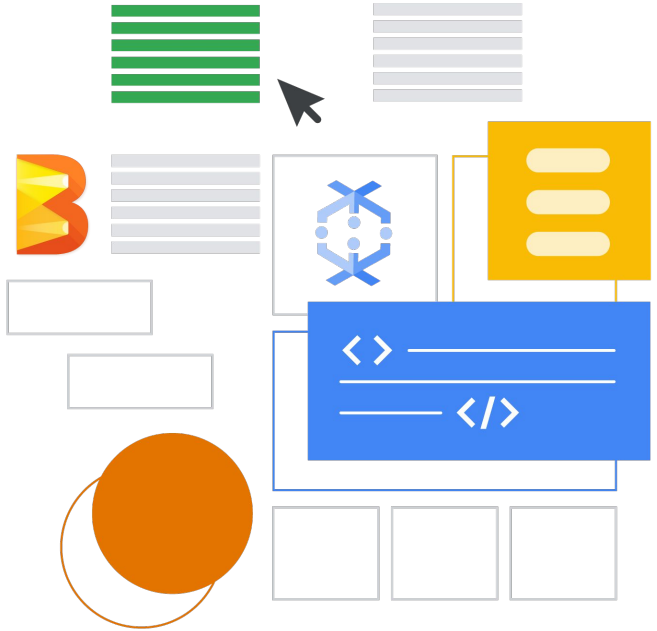
Portability framework

- **Language-agnostic** for representation and **protocol** for execution



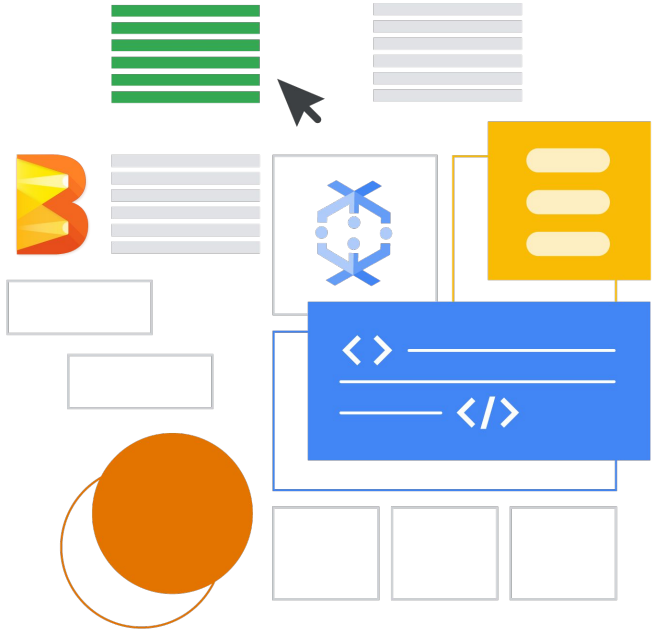
Portability framework

- **Language-agnostic** for representation and **protocol** for execution
- Interoperability layer = **Portability API**

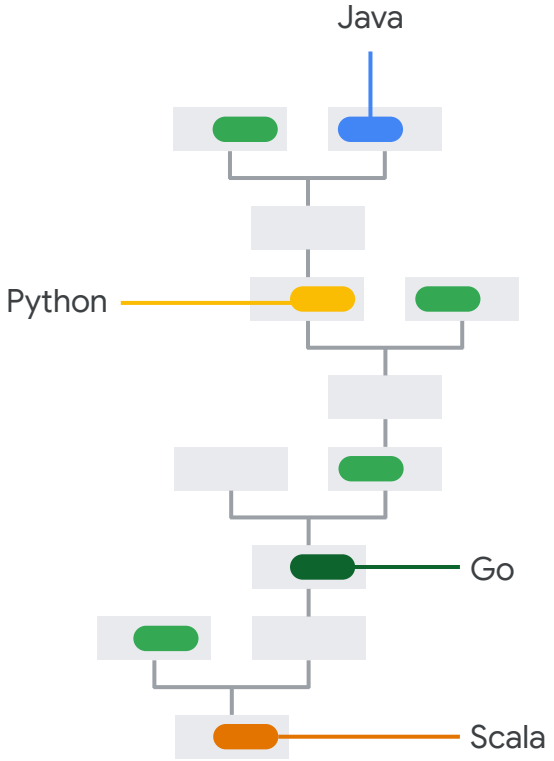


Portability framework

- **Language-agnostic** for representation and **protocol** for execution
- Interoperability layer = **Portability API**
- **Docker containerization** to customize your execution environment

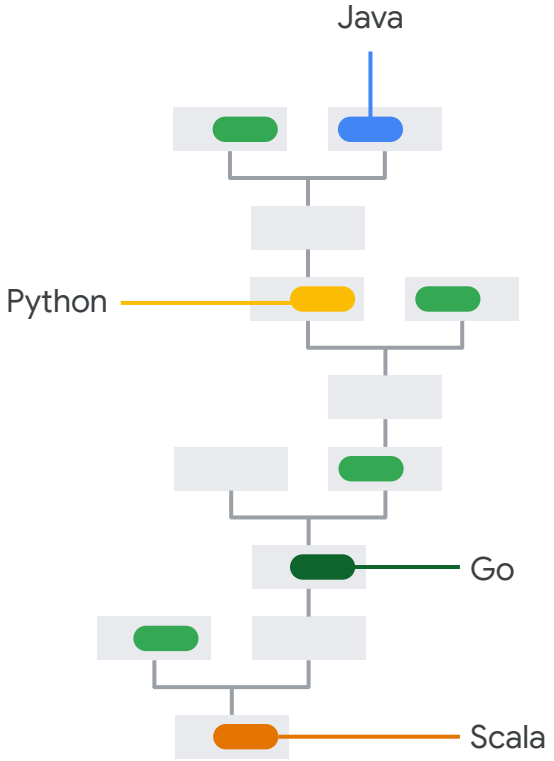


Benefits of portability



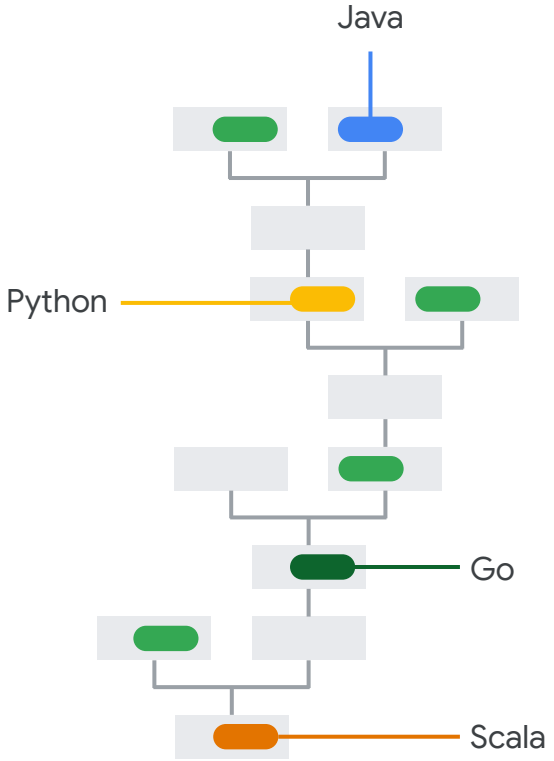
Benefits of portability

- Every runner works with every language.



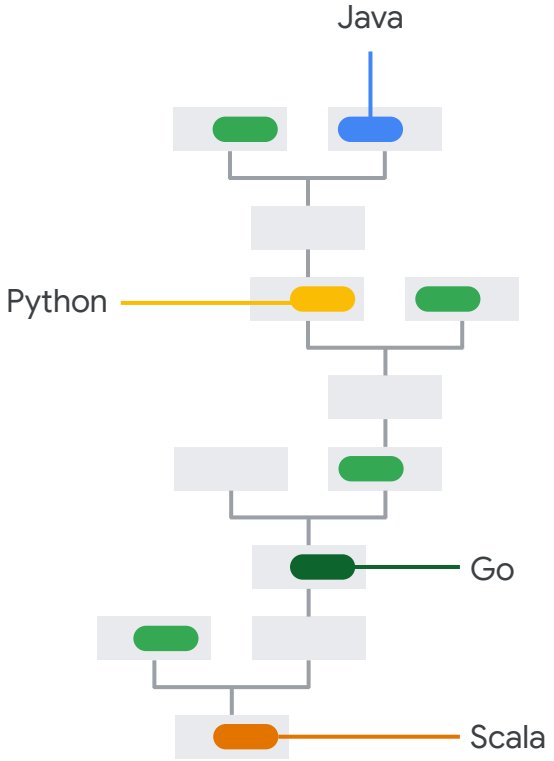
Benefits of portability

- Every runner works with every language
- Configurable, hermetic worker environment



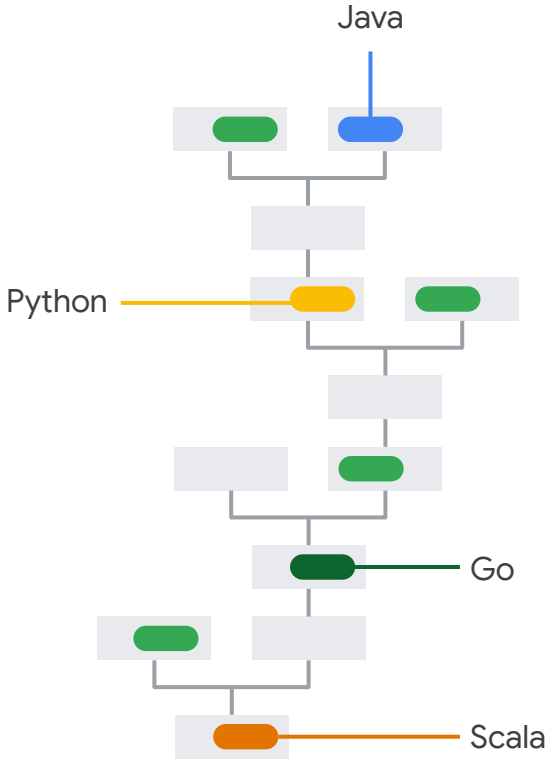
Benefits of portability

- Every runner works with every language
- Configurable, hermetic worker environment
- Multi-language pipelines



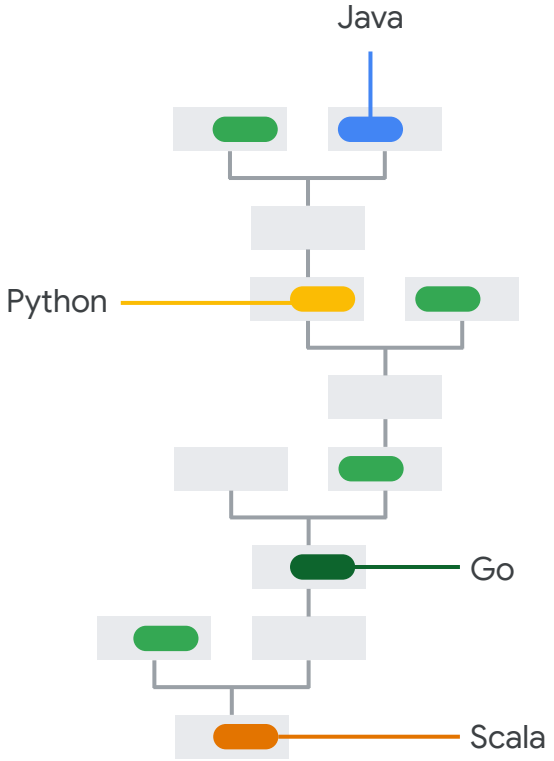
Benefits of portability

- **Every runner** works with **every language**
- Configurable, hermetic **worker environment**
- **Multi-language** pipelines
- **Cross-language** transforms



Benefits of portability

- **Every runner** works with **every language**
- Configurable, hermetic **worker environment**
- **Multi-language** pipelines
- **Cross-language** transforms
- Faster delivery of **new features**



Multi-language pipelines

Agenda

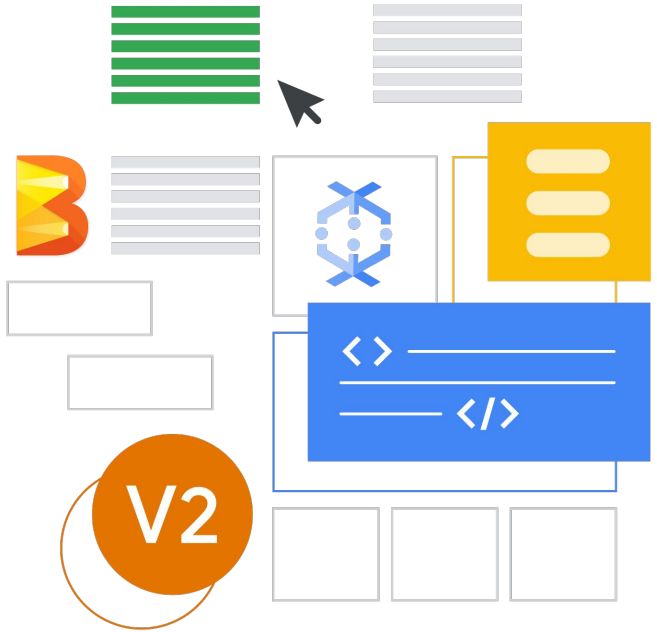
Beam portability

Runner v2

Cross-language transforms

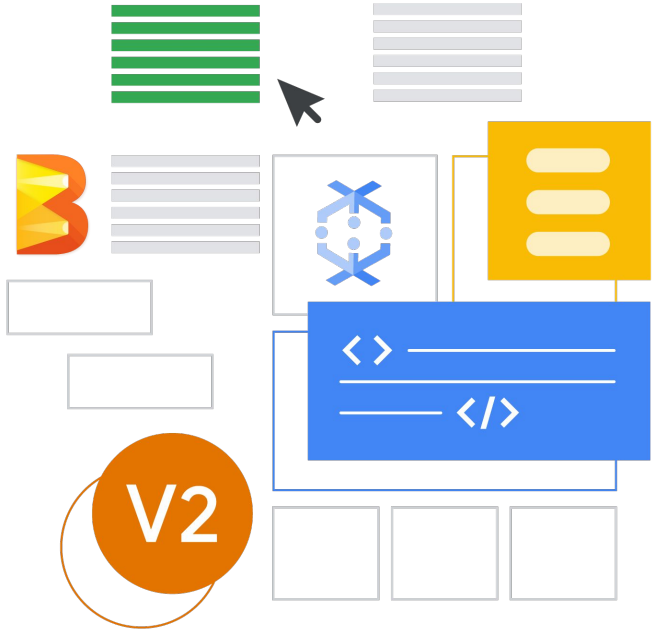
Dataflow Runner v2

- More efficient and portable worker architecture
- Based on Apache Beam portability framework



Dataflow Runner v2

- More efficient and portable worker architecture
- Based on Apache Beam portability framework
- Support for multi-language pipelines and custom containers

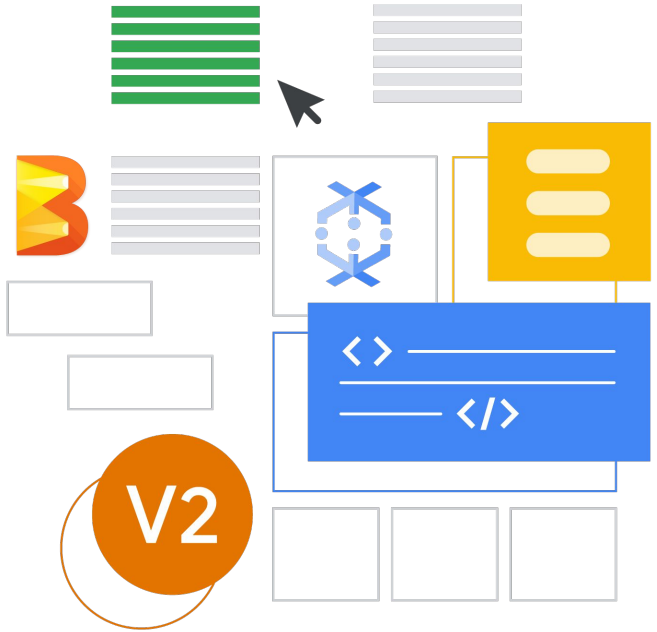


Dataflow Runner v2

- More efficient and portable worker architecture
- Based on Apache Beam portability framework
- Support for multi-language pipelines and custom containers

To enable Runner v2, refer to Dataflow official documentation:

https://cloud.google.com/dataflow/docs

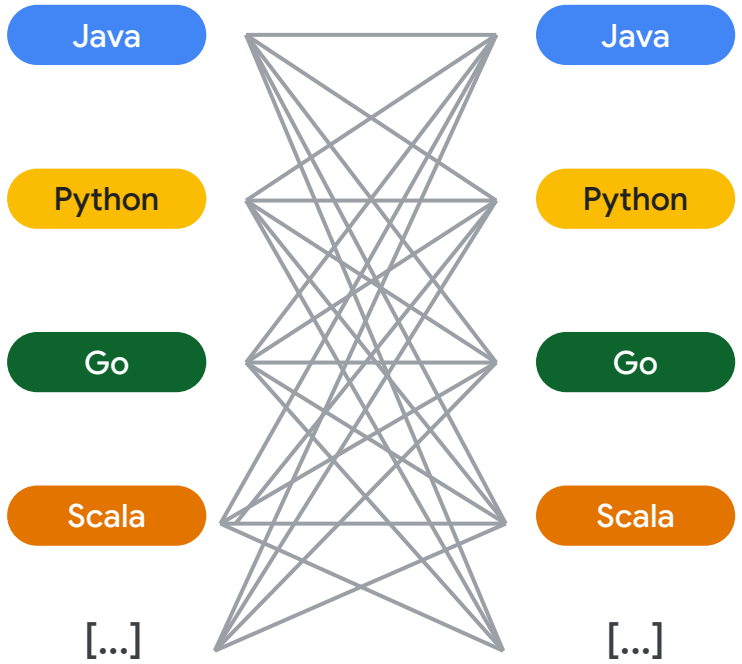


Multi-language pipelines

Agenda

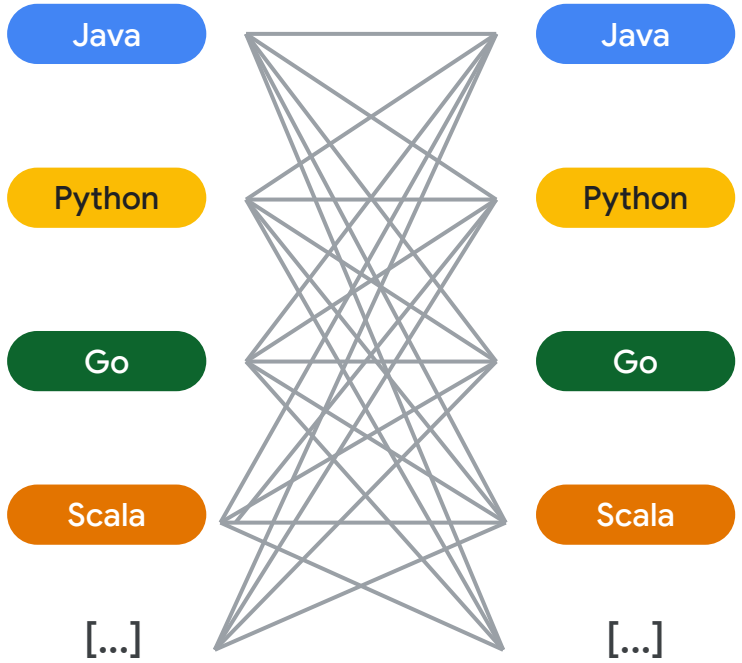
Cross-language transforms

- Transforms can be **shared** among SDKs.



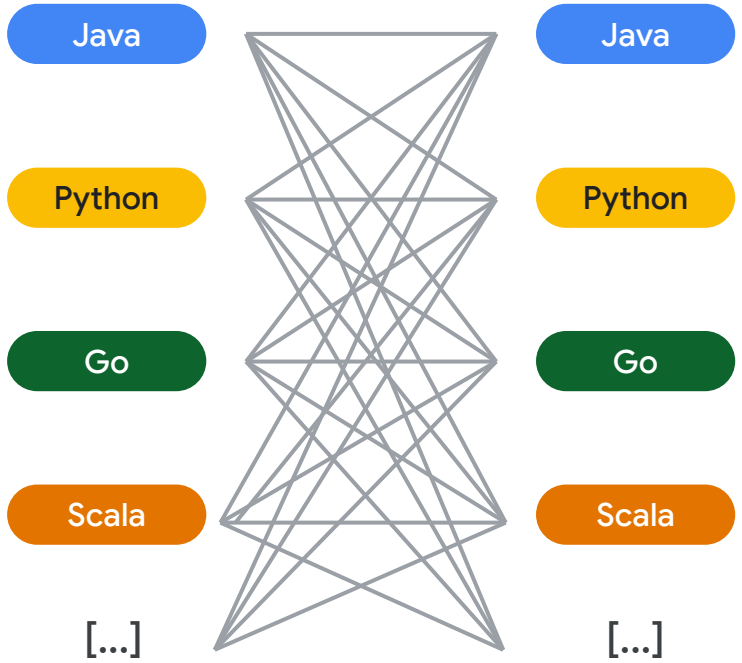
Cross-language transforms

- Transforms can be **shared** among SDKs.
- A **rich set of IOs** from Java is available everywhere.



Cross-language transforms

- Transforms can be **shared** among SDKs.
- A **rich set of IOs** from Java is available everywhere.
- More libraries are available in the **language of your choice**.

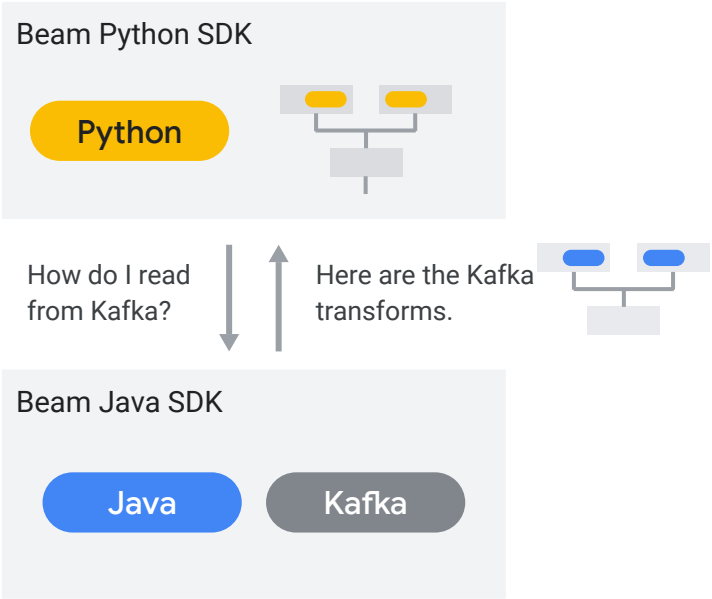


Example of cross-language transform

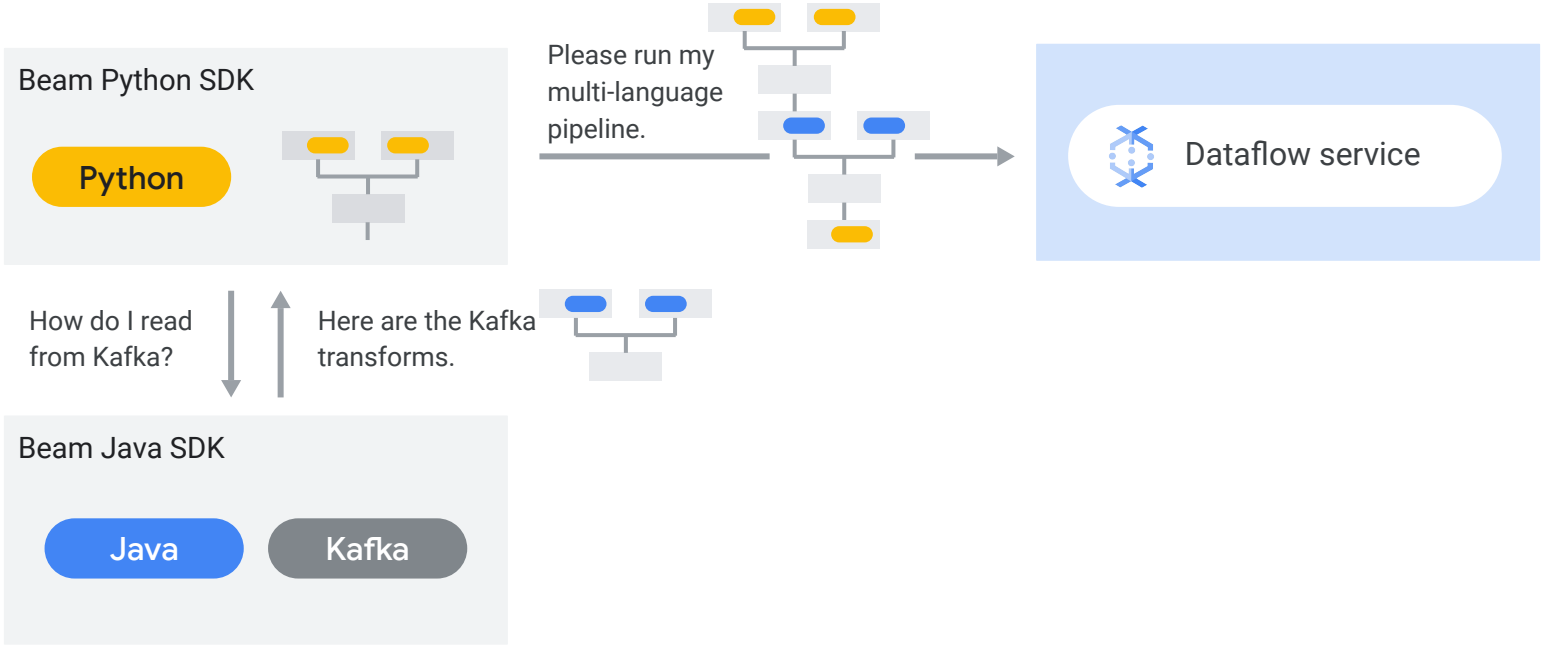
```
from apache_beam.io.kafka import ReadFromKafka

with beam.Pipeline(options=<Your Beam PipelineOptions object>) as p:
      p
      | ReadFromKafka(
          consumer_config={'bootstrap.servers': ‘<Kafka bootstrap servers list>’},
            topics=[<List of Kafka topics>])
```

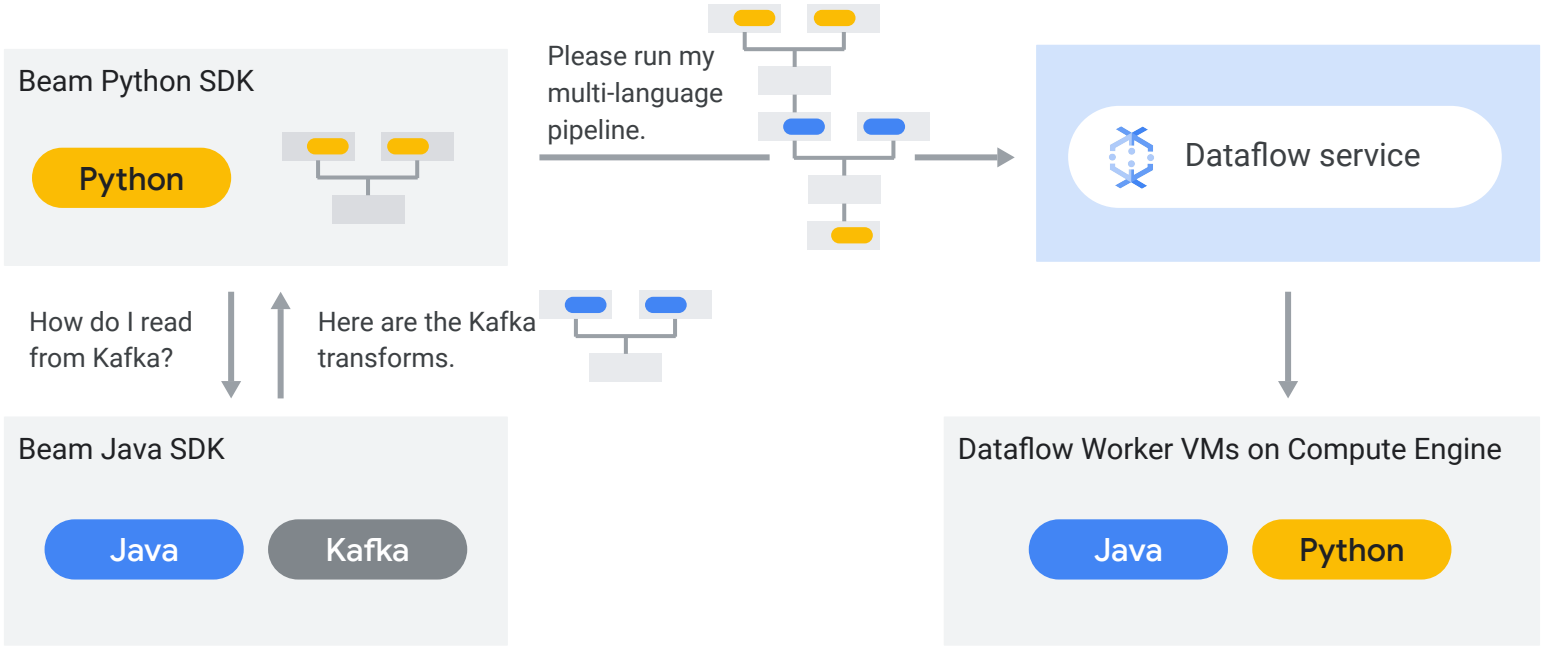
Cross-language transforms



Cross-language transforms



Cross-language transforms



Demo

https://github.com/griscz/beam-college/tree/main/day3/multilanguage-pipelines-demo

Thank you!

References:

https://beam.apache.org/roadmap/connectors-multi-sdk/

https://cloud.google.com/blog/products/data-analytics/multi-language-sdks-for-building-cloud-pipelines