Separating Compute and Storage with Dataflow

Federico Patota

Cloud Consultant, Google Cloud

Agenda

Course Intro

Beam and Dataflow Refresher

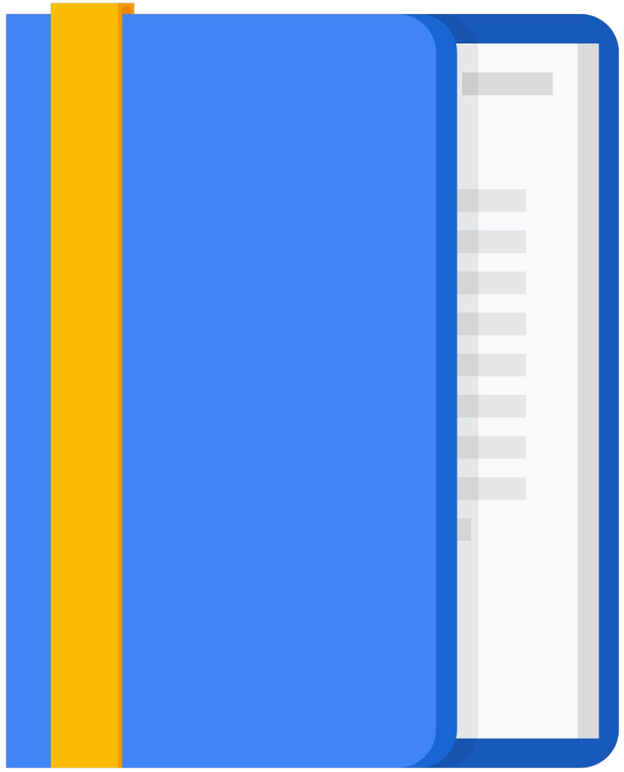
Beam Portability

Separating Compute and Storage

IAM, Quotas, and Permissions

Security

Summary

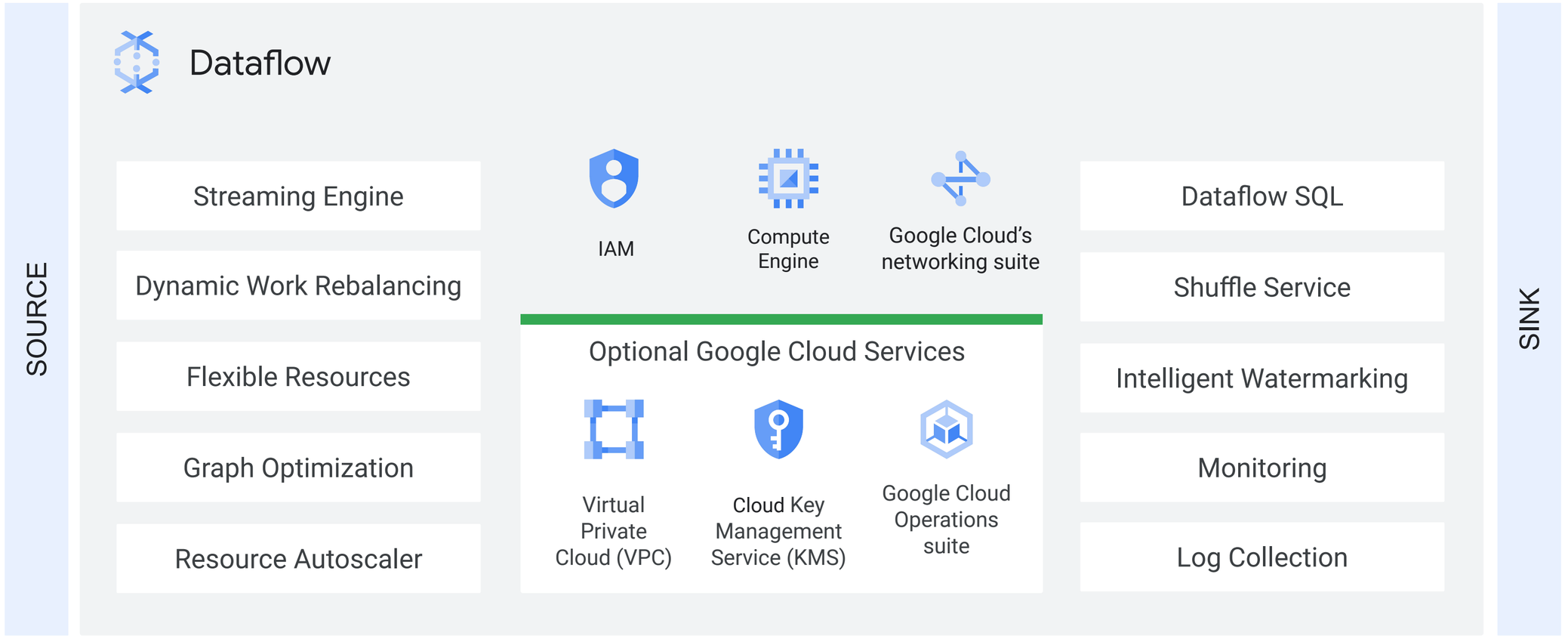


Separating compute and storage with Dataflow

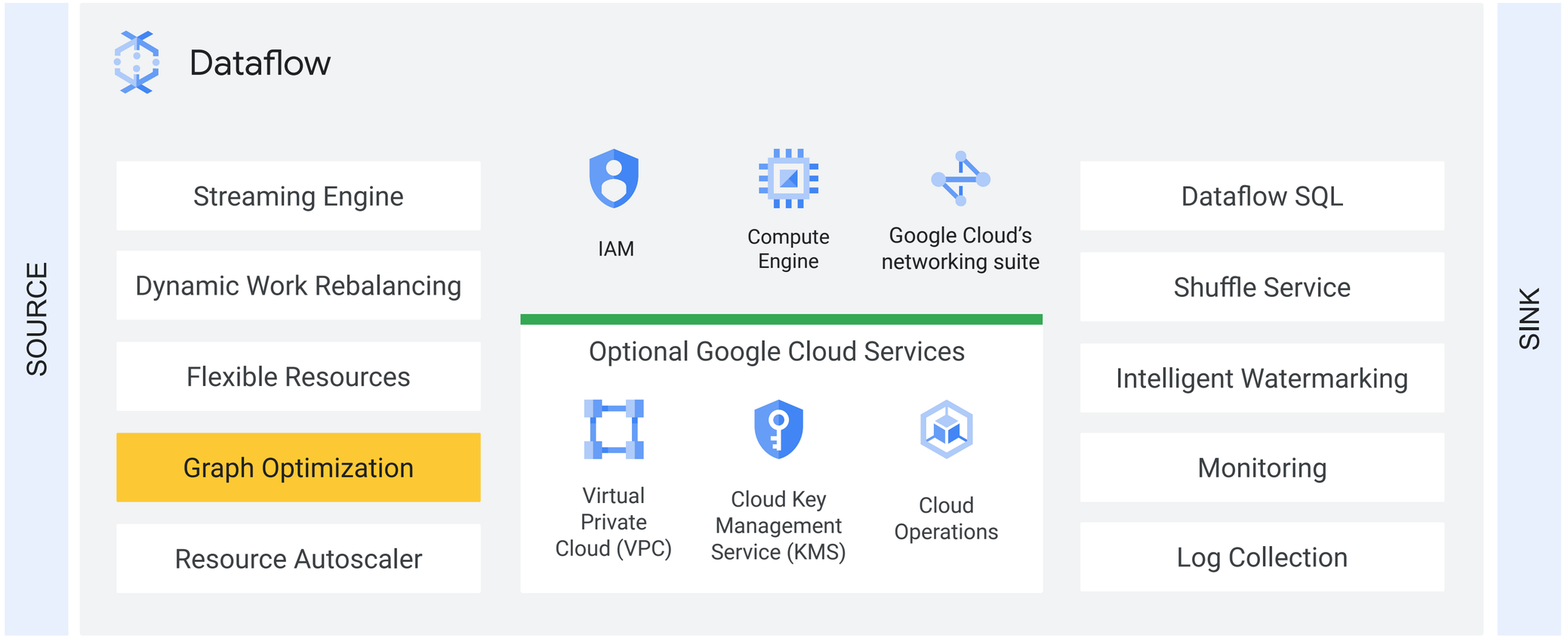
Agenda

- Dataflow
- Dataflow Shuffle Service
- Dataflow Streaming Engine
- Flexible Resource Scheduling

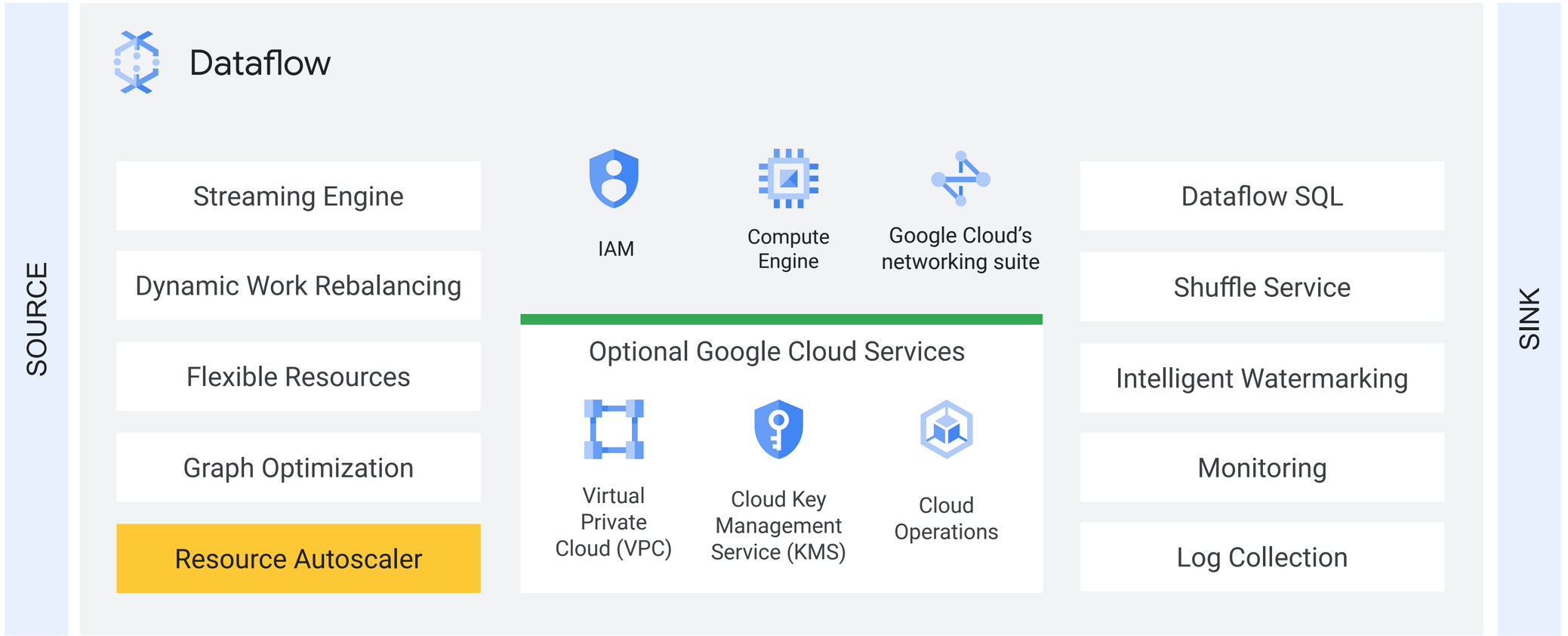
The Google Cloud runner: Dataflow



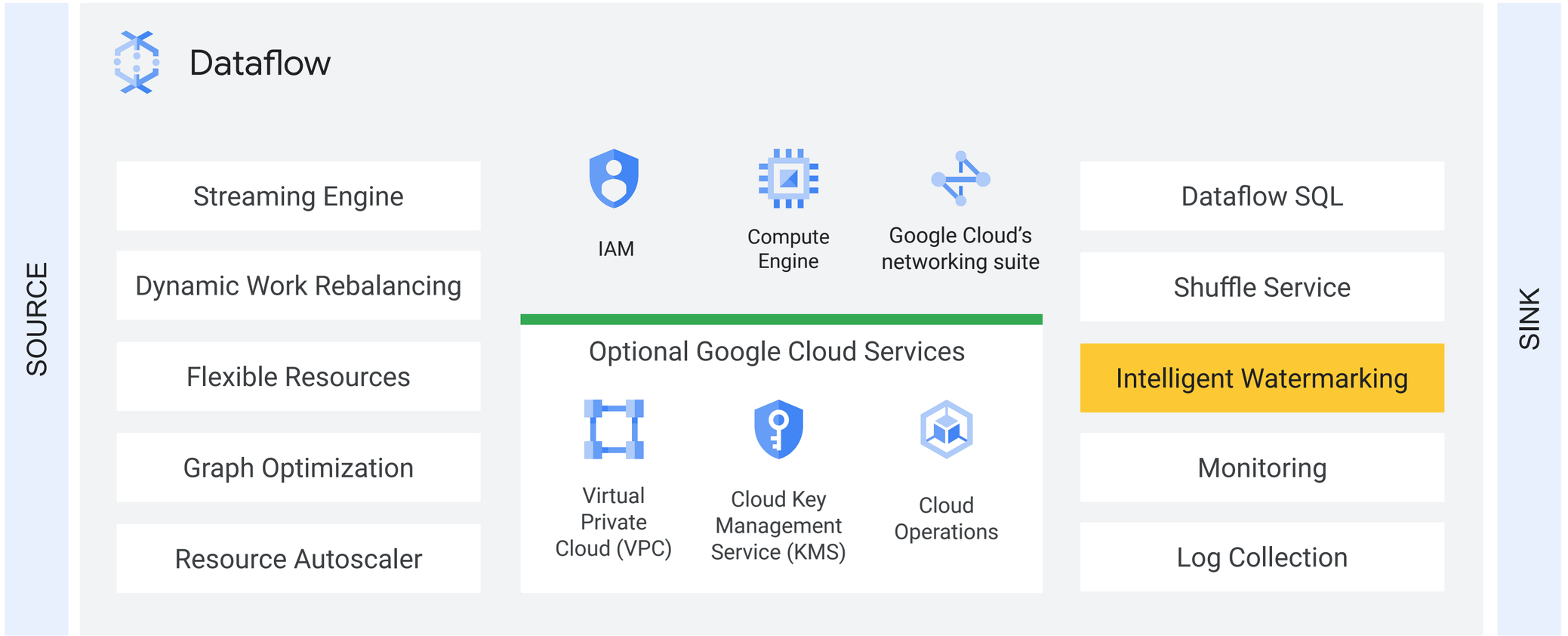
The Google Cloud runner: Dataflow



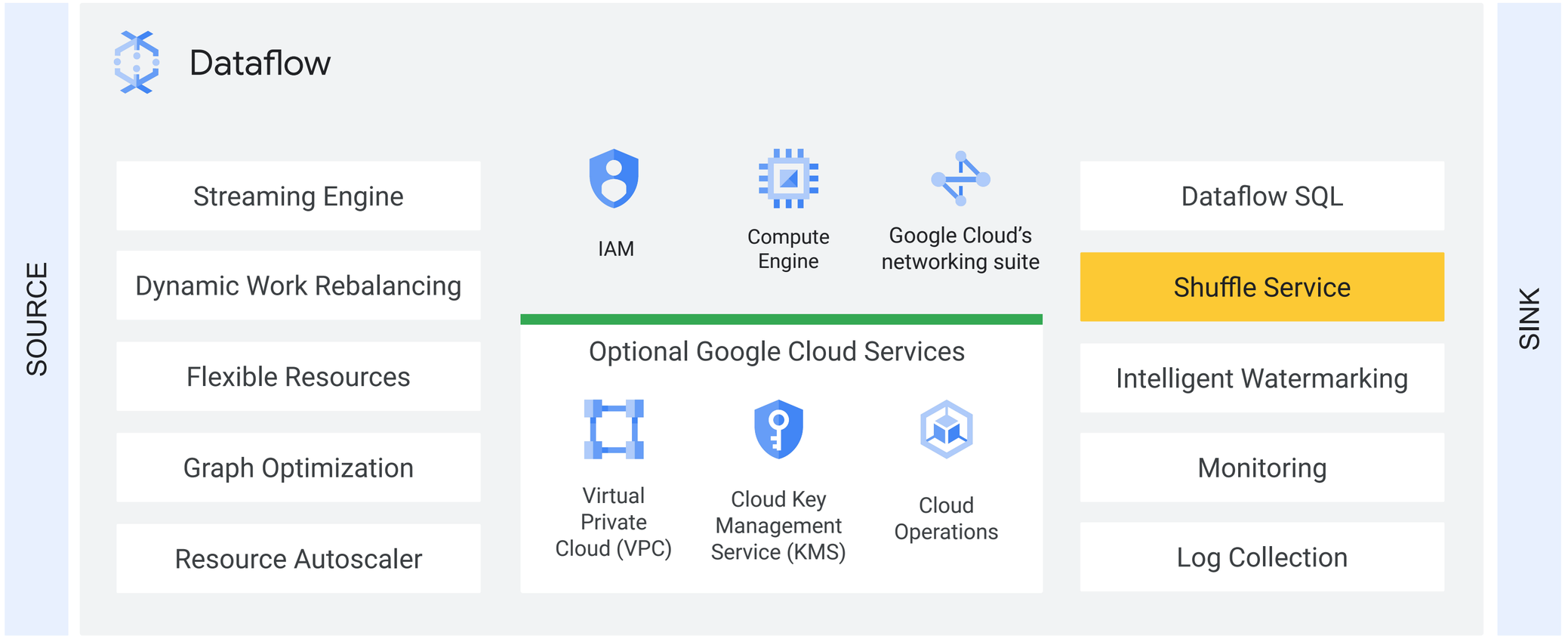
The Google Cloud runner: Dataflow



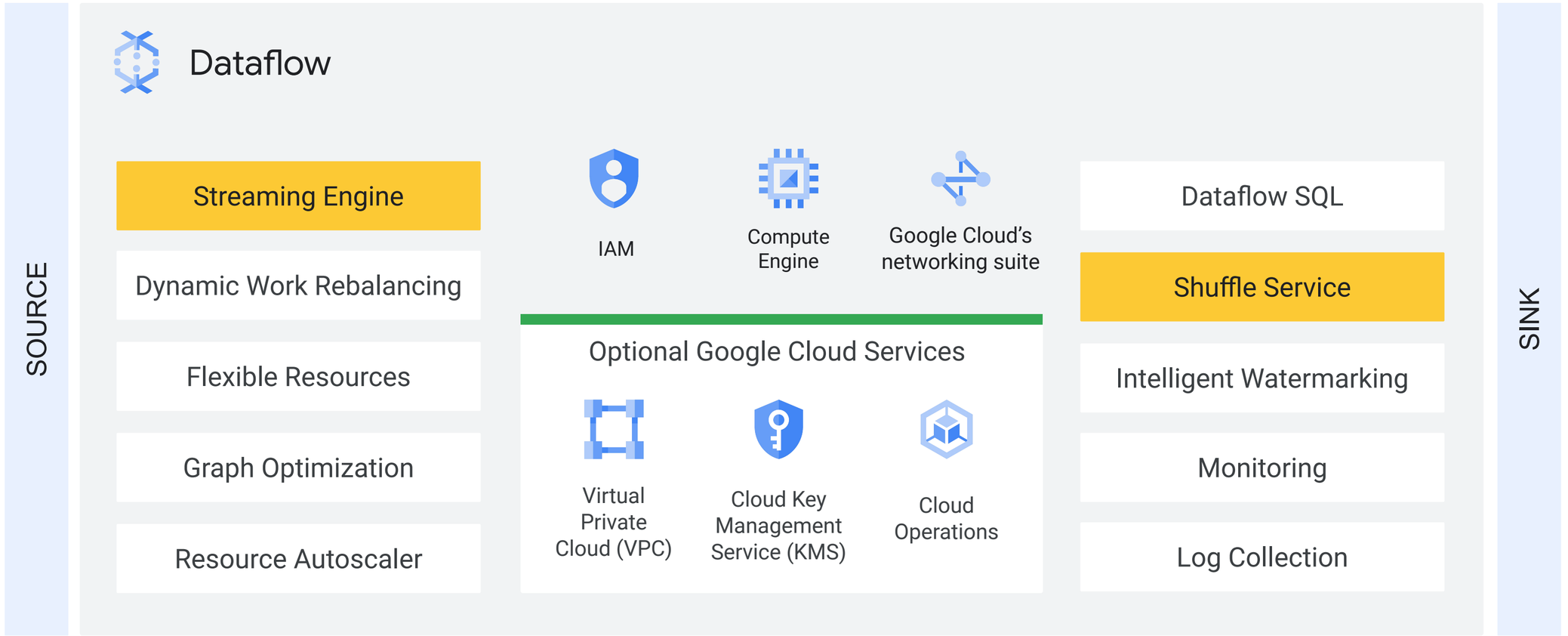
The Google Cloud runner: Dataflow



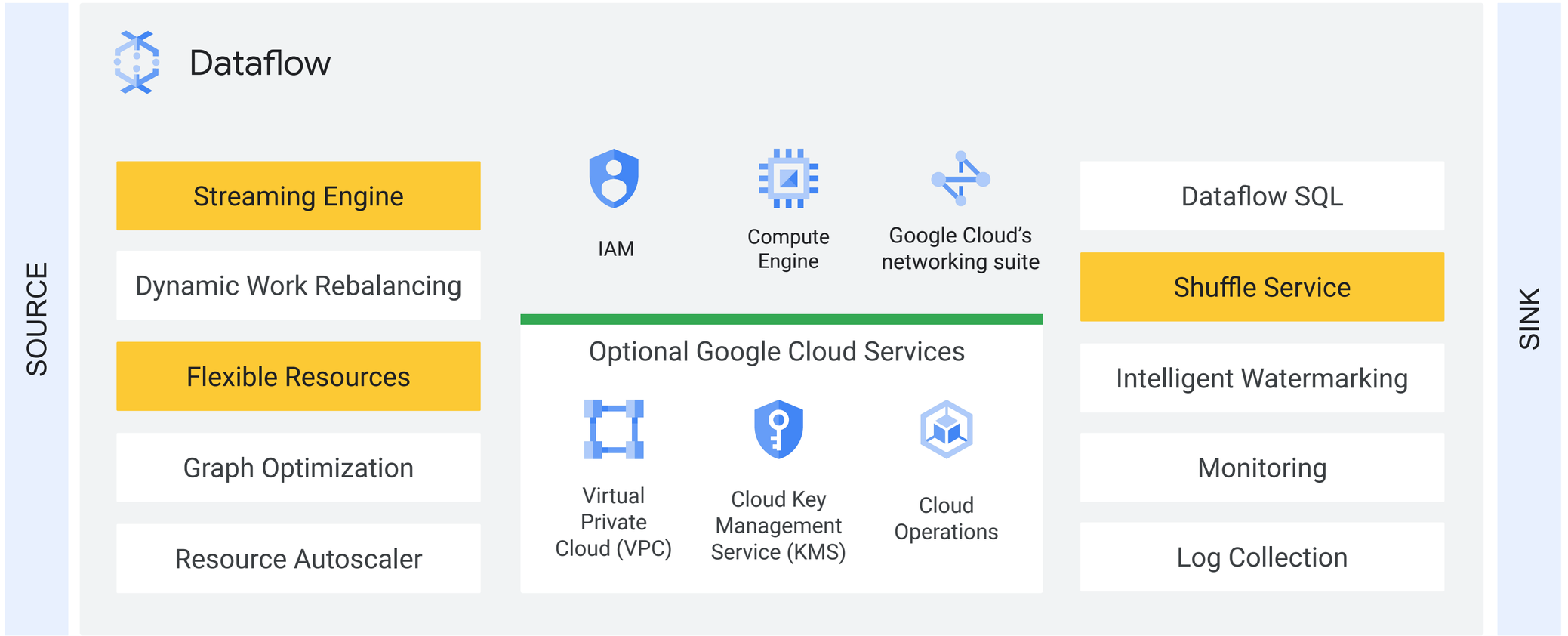
The Google Cloud runner: Dataflow



The Google Cloud runner: Dataflow



The Google Cloud runner: Dataflow



Separating compute and storage with Dataflow

Agenda

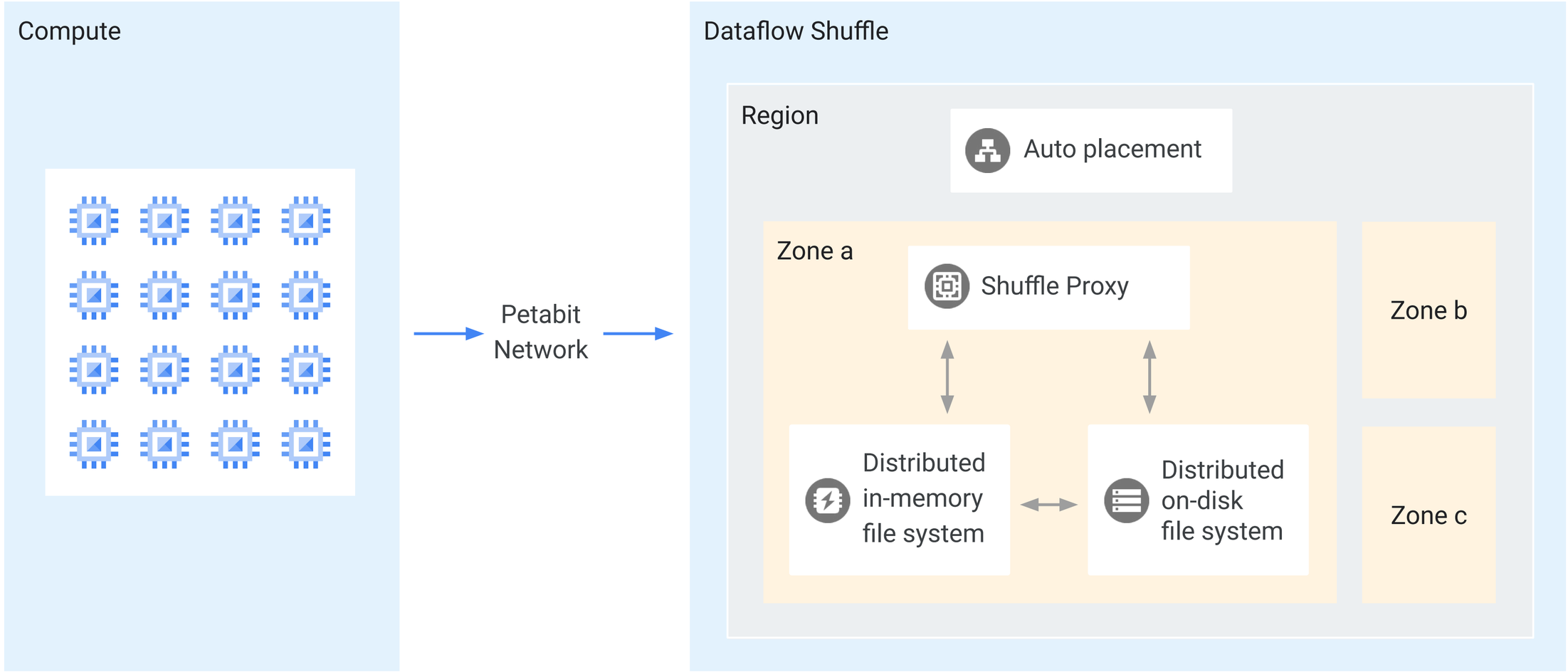
Dataflow

Dataflow Shuffle Service

Dataflow Streaming Engine

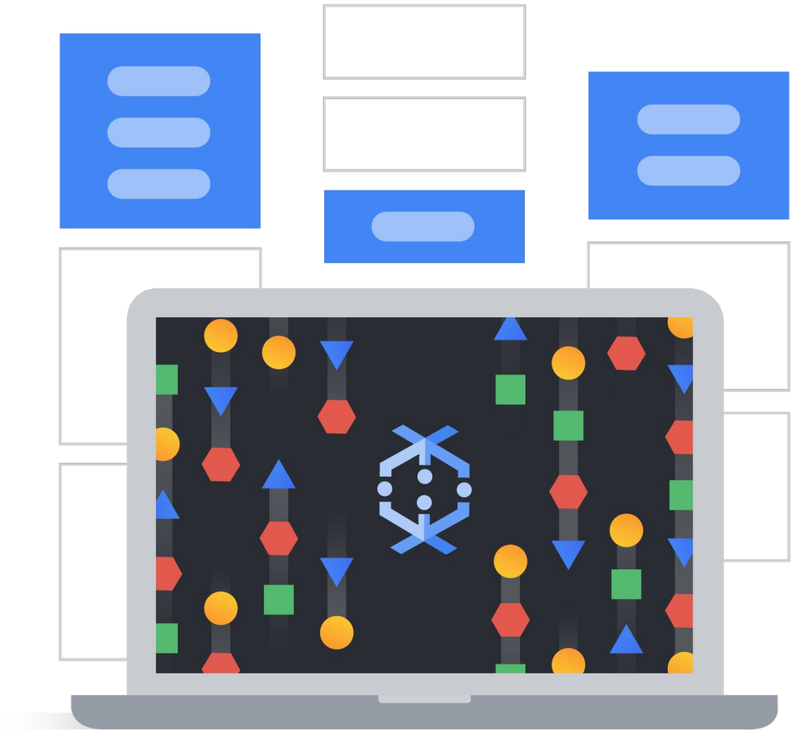
Flexible Resource Scheduling

Dataflow Shuffle service: Batch



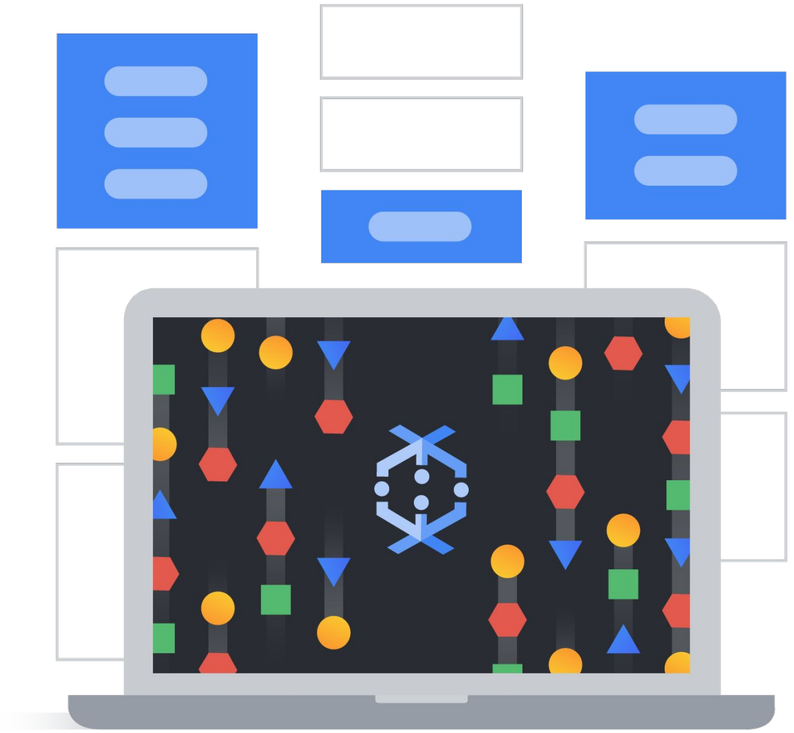
Dataflow Shuffle service: Batch

- **Faster execution time** of batch pipelines for most cases



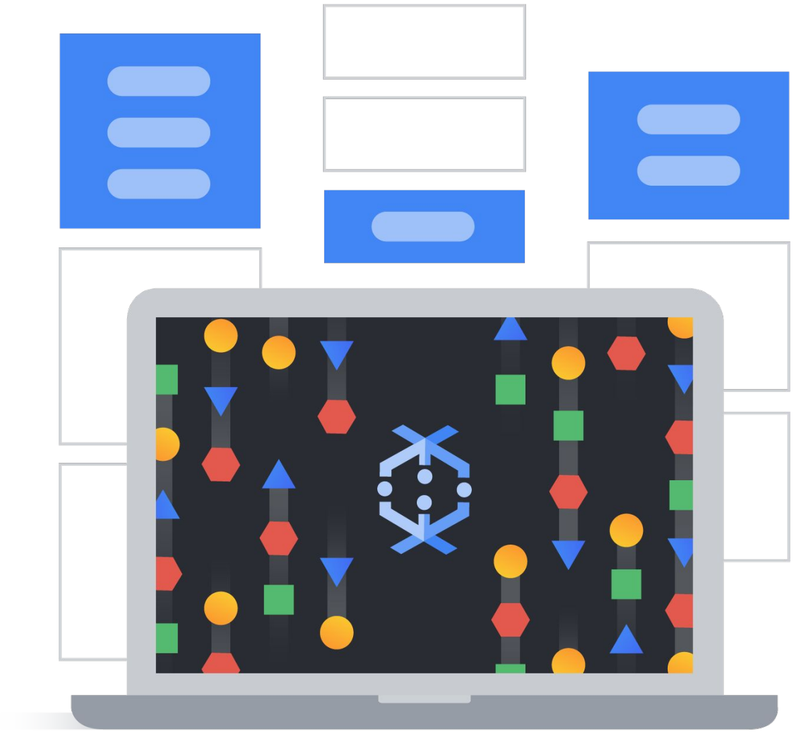
Dataflow Shuffle service: Batch

- **Faster execution time** of batch pipelines for most cases
- Reduced consumption of worker's **CPU**, **memory**, and **storage**



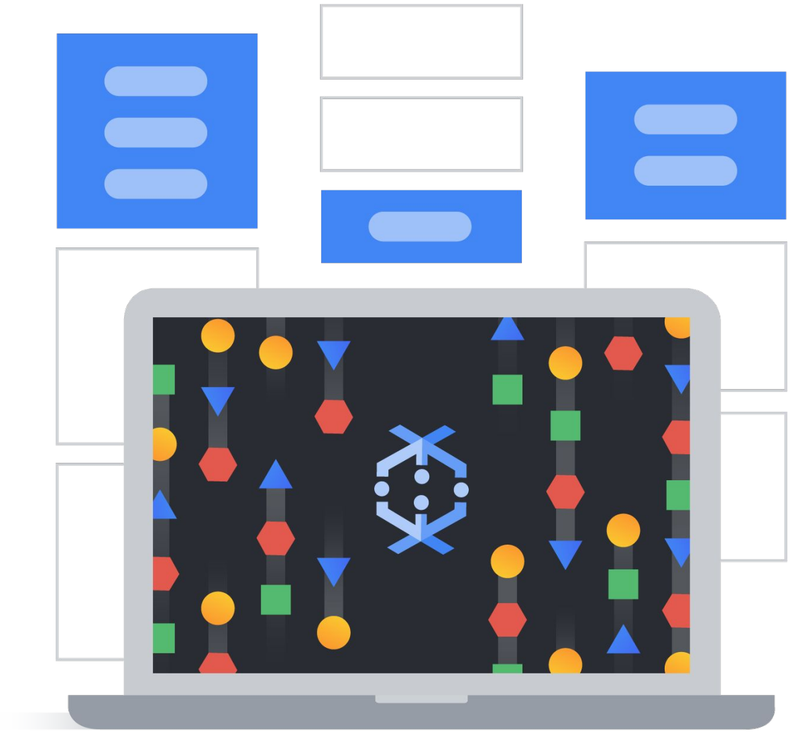
Dataflow Shuffle service: Batch

- **Faster execution time** of batch pipelines for most cases
- Reduced consumption of worker's **CPU**, **memory**, and **storage**
- Better **autoscaling**



Dataflow Shuffle service: Batch

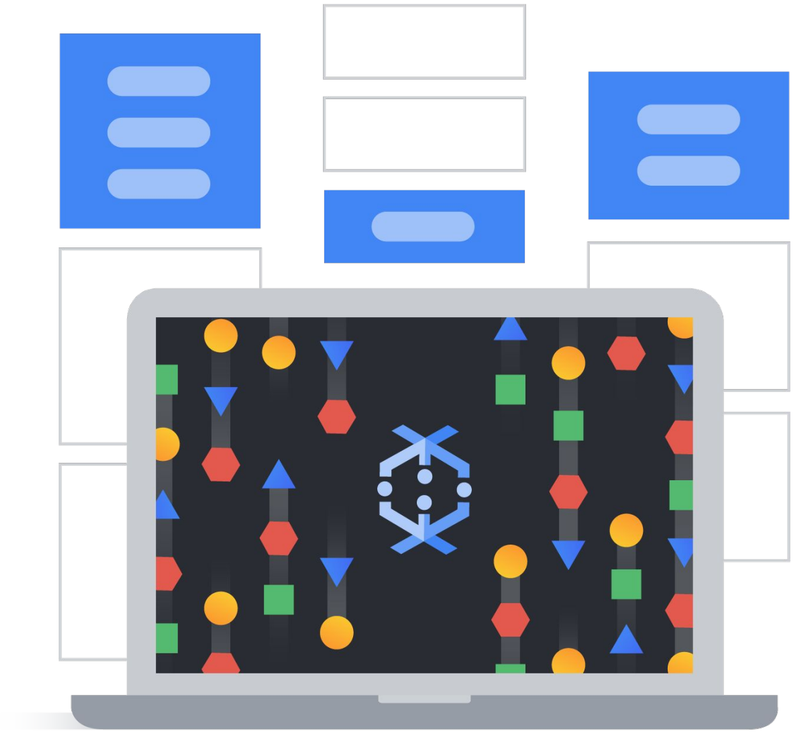
- **Faster execution time** of batch pipelines for most cases
- Reduced consumption of worker's **CPU**, **memory**, and **storage**
- Better **autoscaling**
- Better **fault tolerance**



Dataflow Shuffle service: Batch

- **Faster execution time** of batch pipelines for most cases
- Reduced consumption of worker's **CPU**, **memory**, and **storage**
- Better **autoscaling**
- Better **fault tolerance**

For information about Dataflow Shuffle service, see the official Dataflow documentation.



Separating compute and storage with Dataflow

Agenda

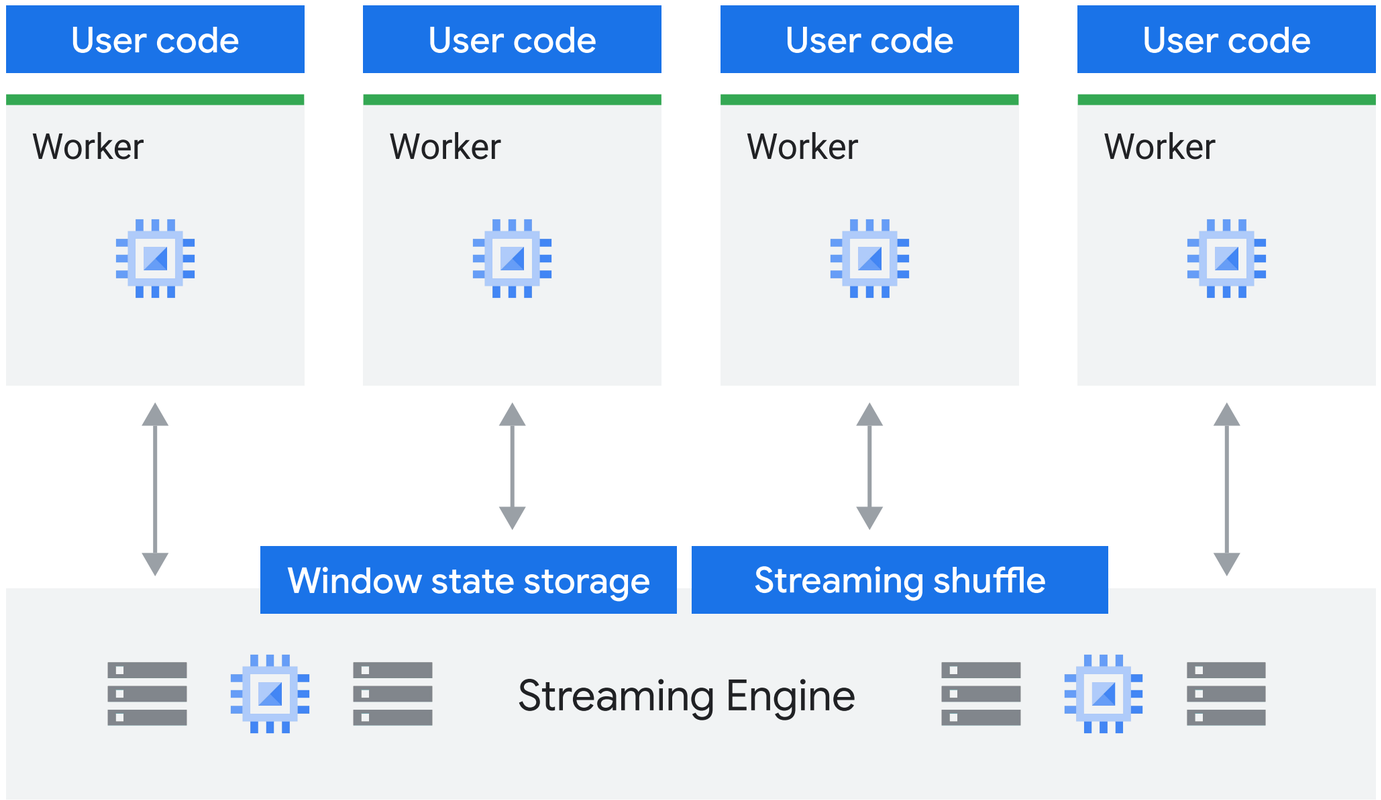
Dataflow

Dataflow Shuffle Service

Dataflow Streaming Engine

Flexible Resource Scheduling

Dataflow Streaming Engine



Dataflow Streaming Engine

- Reduced consumption of worker CPU, memory, and storage

Dataflow Streaming Engine

- Reduced consumption of worker CPU, memory, and storage
- Lower resource and quota consumption

Dataflow Streaming Engine

- Reduced consumption of worker **CPU**, **memory**, and **storage**
- Lower resource and quota consumption
- More **responsive autoscaling** for incoming data volume variations

Dataflow Streaming Engine

- Reduced consumption of worker **CPU**, **memory**, and **storage**
- Lower resource and quota consumption
- More **responsive autoscaling** for incoming data volume variations
- Improved **supportability**

Dataflow Streaming Engine

- Reduced consumption of worker **CPU**, **memory**, and **storage**
- Lower resource and quota consumption
- More **responsive autoscaling** for incoming data volume variations
- Improved **supportability**

For information about Dataflow Streaming Engine, see the official Dataflow documentation.

Separating compute and storage with Dataflow

Agenda

Dataflow

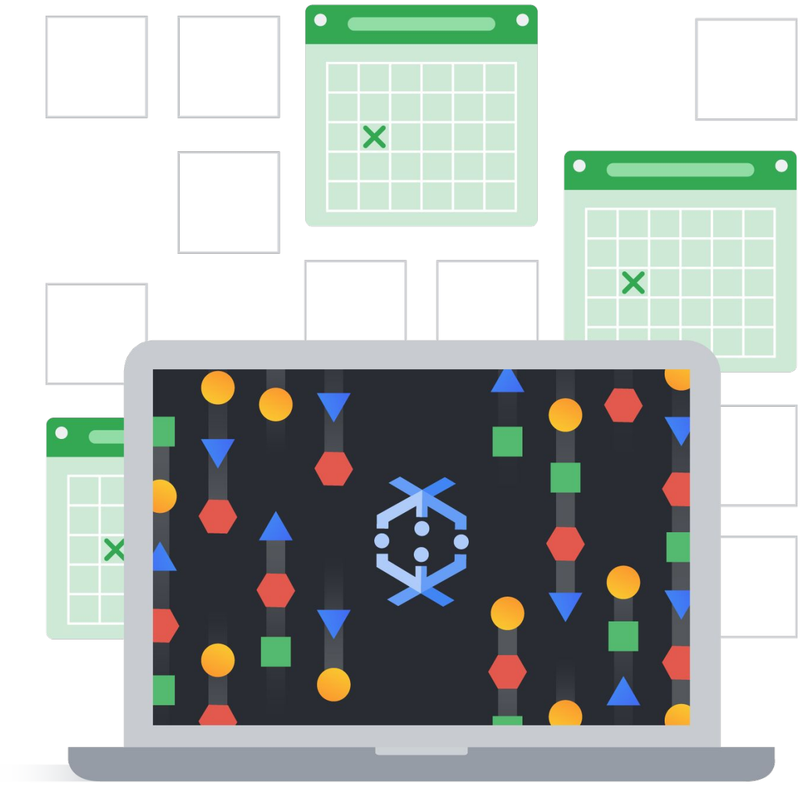
Dataflow Shuffle Service

Dataflow Streaming Engine

Flexible Resource Scheduling

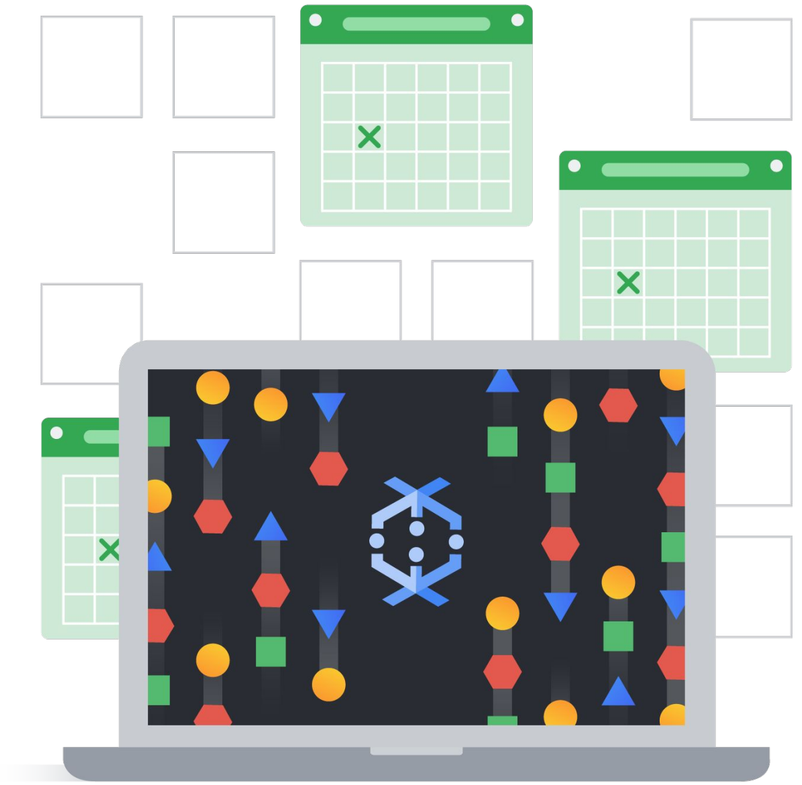
Flexible Resource Scheduling (FlexRS)

- **Reduced** batch processing costs because of:
  - Advanced **scheduling**
  - Dataflow **Shuffle** Service
  - Mix of **preemptible** and **standard** VMs



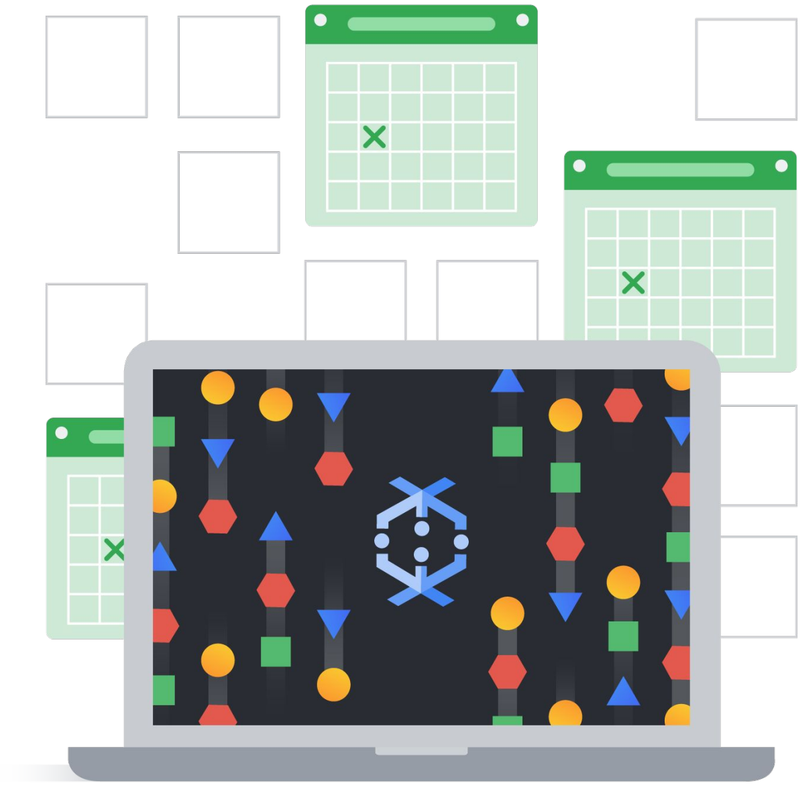
Flexible Resource Scheduling (FlexRS)

- **Reduced** batch processing costs because of:
  - Advanced **scheduling**
  - Dataflow **Shuffle** Service
  - Mix of **preemptible** and **standard** VMs
- **Execution** within 6 hours from job creation



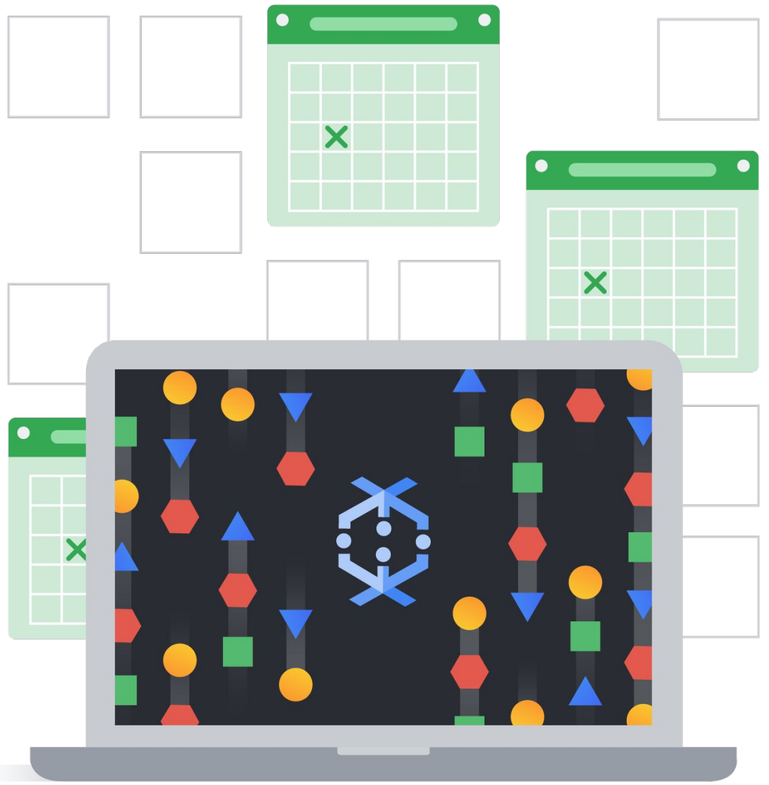
Flexible Resource Scheduling (FlexRS)

- **Reduced** batch processing costs because of:
  - Advanced **scheduling**
  - Dataflow **Shuffle** Service
  - Mix of **preemptible** and **standard** VMs
- **Execution** within 6 hours from job creation
- Suitable for workloads that are **not time-critical**



Flexible Resource Scheduling (FlexRS)

- **Reduced** batch processing costs because of:
  - Advanced **scheduling**
  - Dataflow **Shuffle** Service
  - Mix of **preemptible** and **standard** VMs
- **Execution** within 6 hours from job creation
- Suitable for workloads that are **not time-critical**
- Early validation run at job submission



Flexible Resource Scheduling (FlexRS)

- **Reduced** batch processing costs because of:
  - Advanced **scheduling**
  - Dataflow **Shuffle** Service
  - Mix of **preemptible** and **standard** VMs
- **Execution** within 6 hours from job creation
- Suitable for workloads that are **not time-critical**
- Early validation run at job submission

For information about FlexRS, see the official Dataflow documentation.