Creating and Using PCollections and Side Inputs

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Pipeline as directed acyclic graphs (DAGs)

PCollections as edges, transformations as nodes

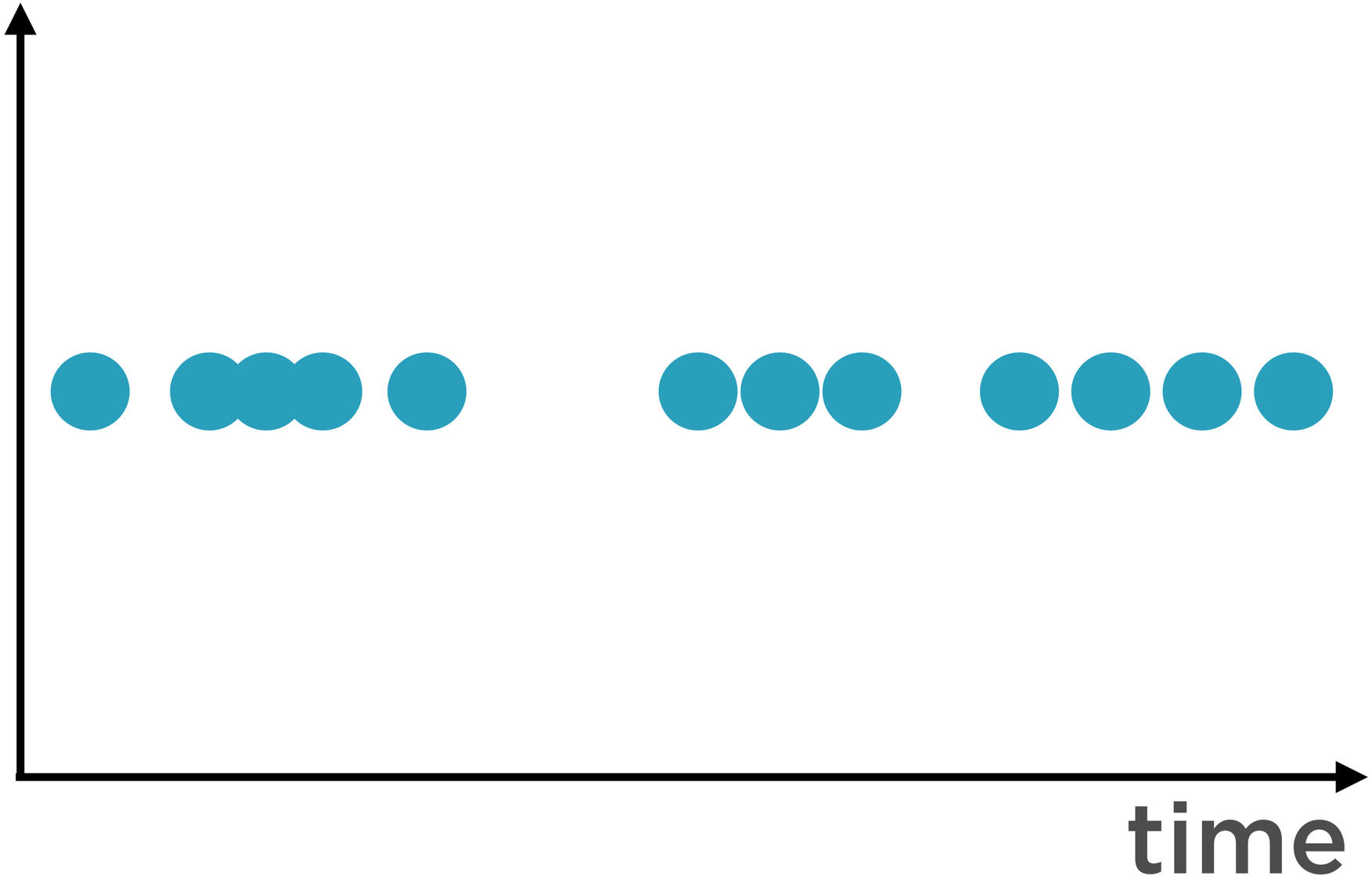
Windowing operations

Branching operations

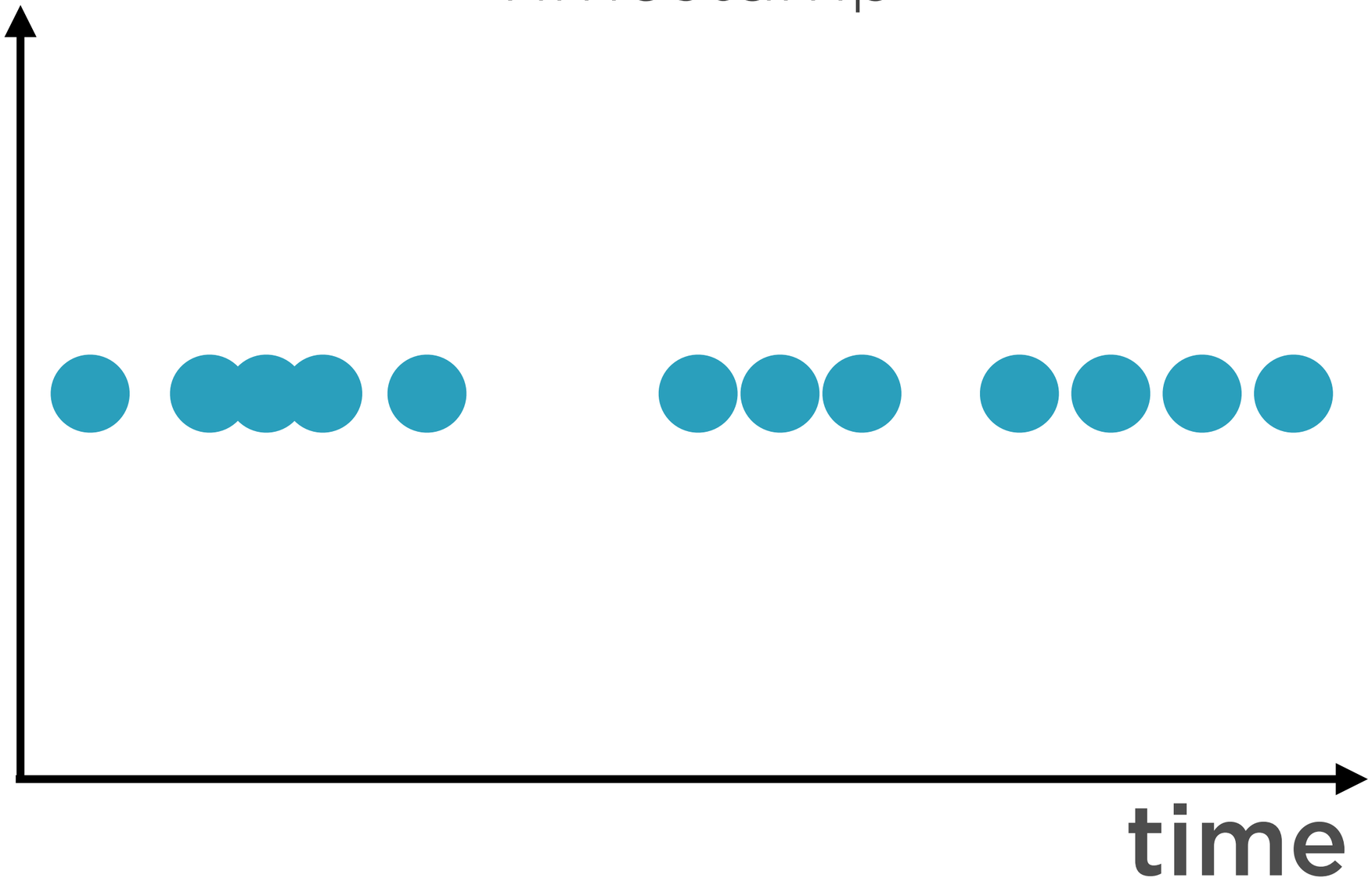
Side inputs into pipelines

Working with Subsets of Data

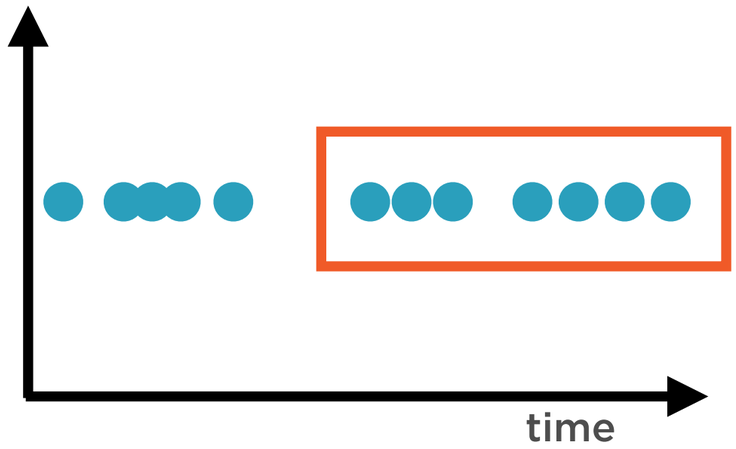
Data Associated with Time



Streaming or Batch Data with Timestamp



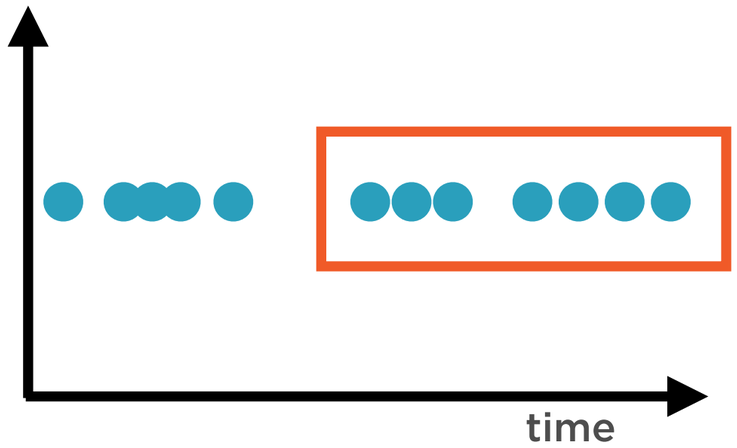
Window Transformations



A window is a subset of a stream based on

- Time interval
- Count of entities
- Interval between entities

Window Transformations



Transformations can be applied on all entities within a window

- sum, min, max, average

Types of Windows

Tumbling Window

Sliding Window

Count Window

Session Window

Global Window

Types of Windows

Tumbling Window

Sliding Window

Count Window

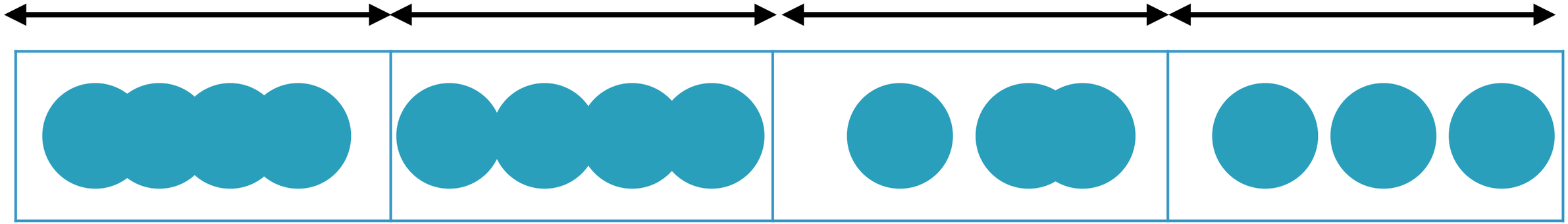
Session Window

Global Window

Types of Windows

A stream of data or batch data with timestamps

Tumbling Window

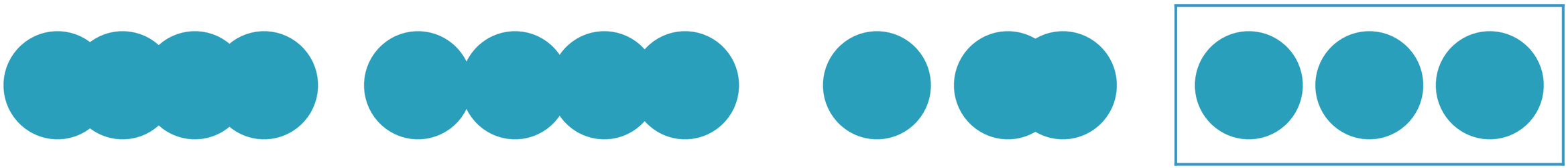


Fixed window size

Non-overlapping time

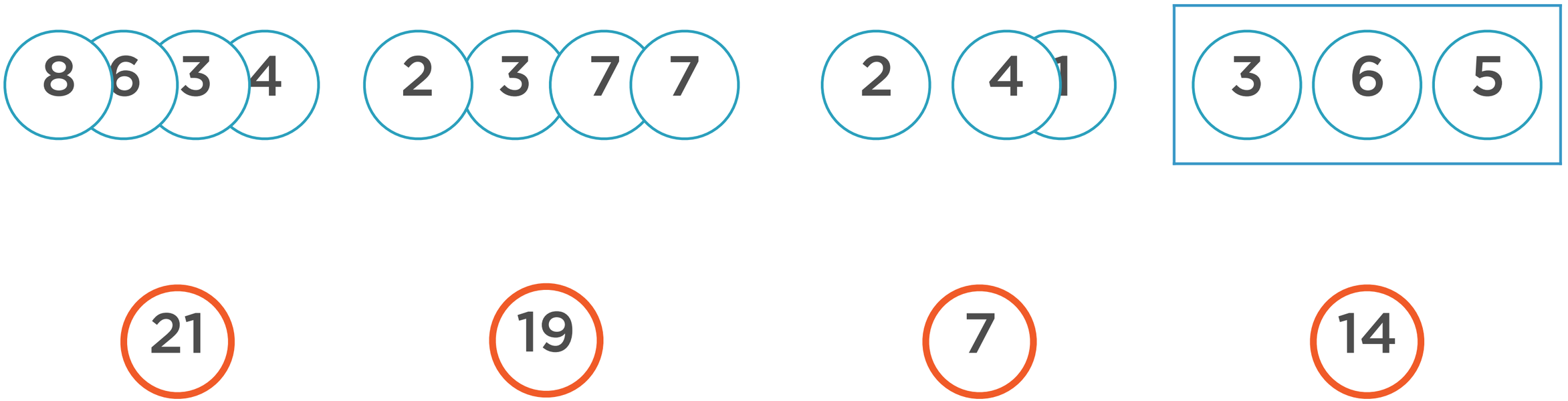
Number of entities differ within a window

Tumbling Window



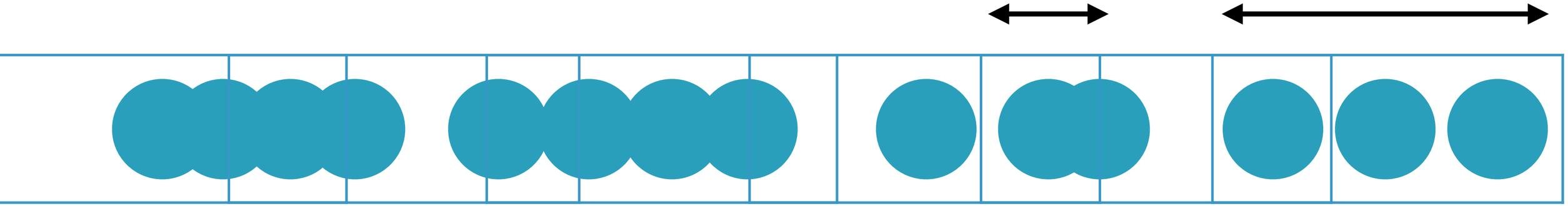
The window tumbles over the data, in a non-overlapping manner

Tumbling Window



Apply the sum() operation on each window

Sliding Window

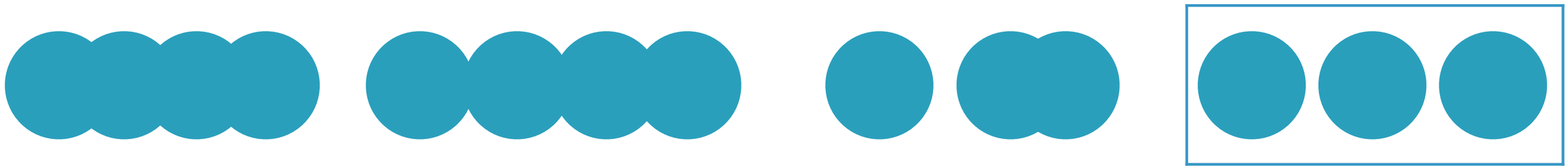


Fixed window size

Overlapping time - sliding interval

Number of entities differ within a window

Sliding Window

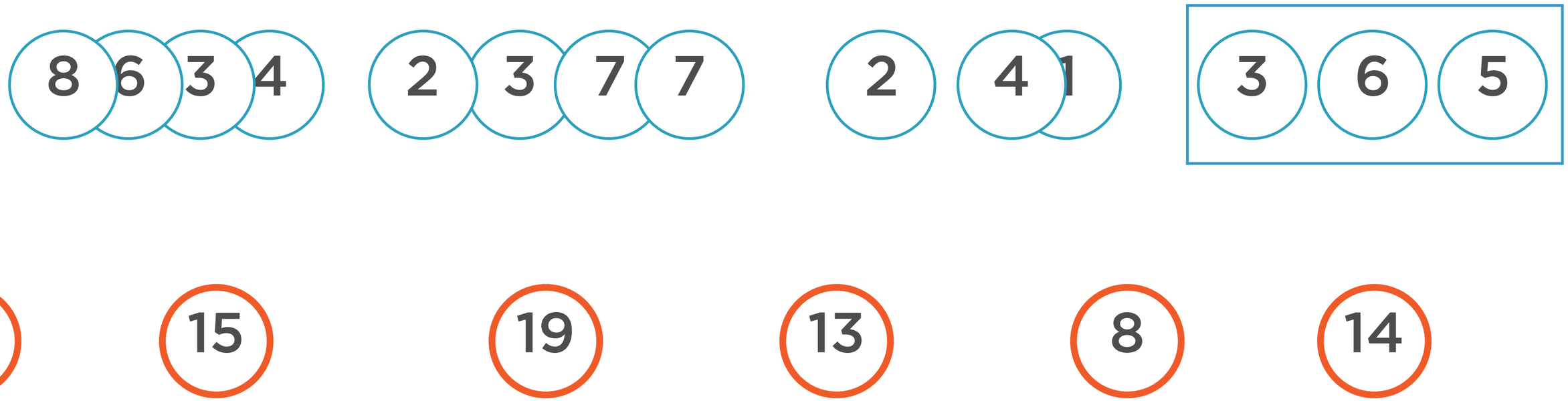


Fixed window size

Overlapping time - sliding interval

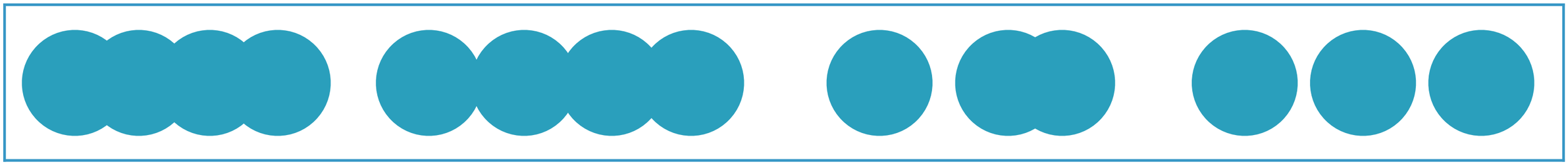
Number of entities differ within a window

Sliding Window



Apply the sum() operation on each window

Global Window



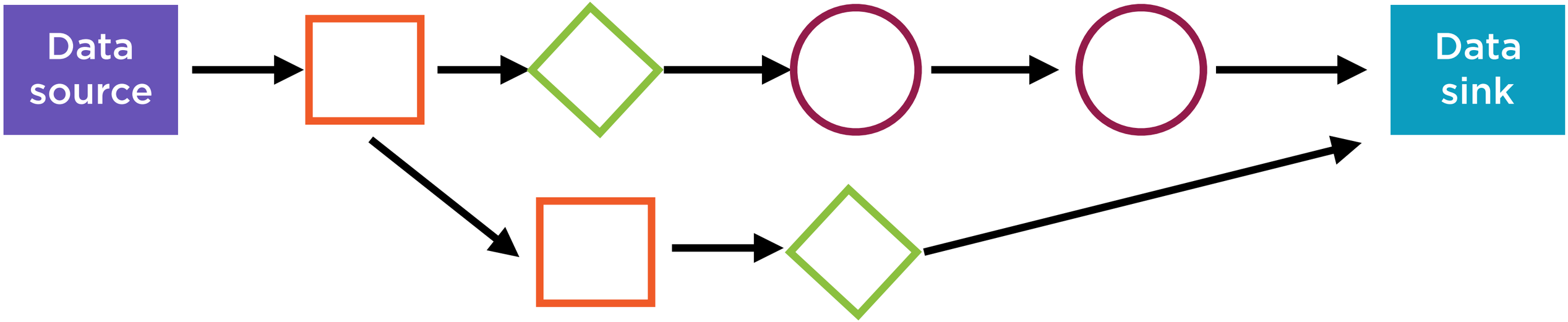
All data in the stream in one window

Demo

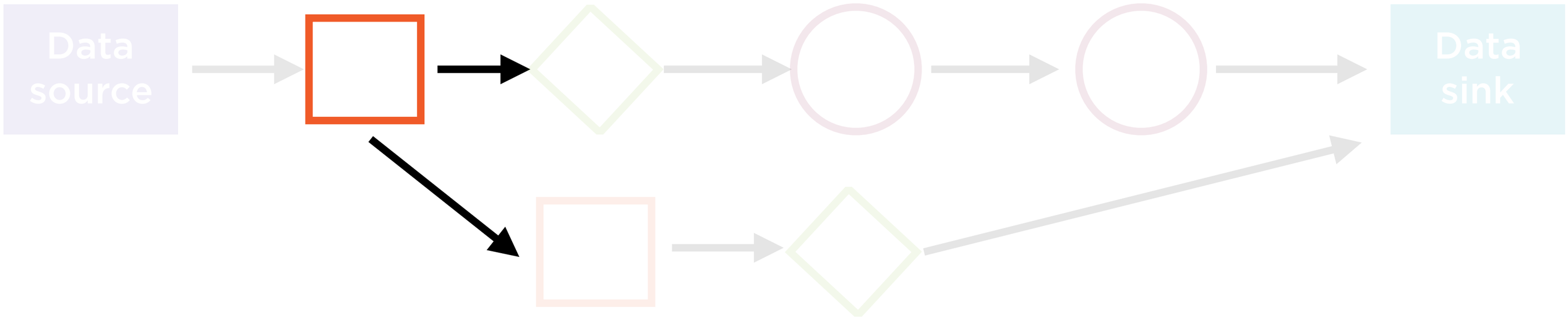
Performing transformations and aggregations using window operations

Branching Operatons

Apache Beam Pipeline

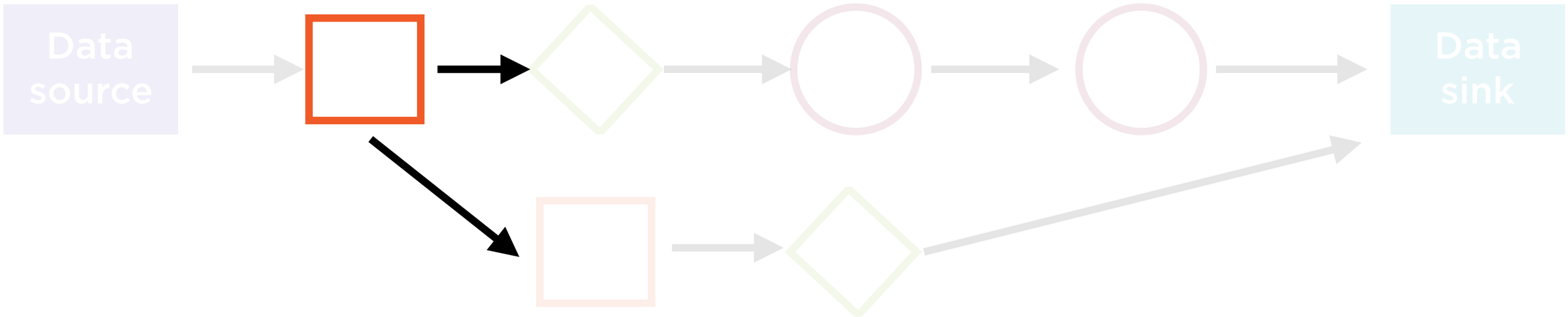


Branching Operations



A single transformation can have multiple outputs

Branching Operations



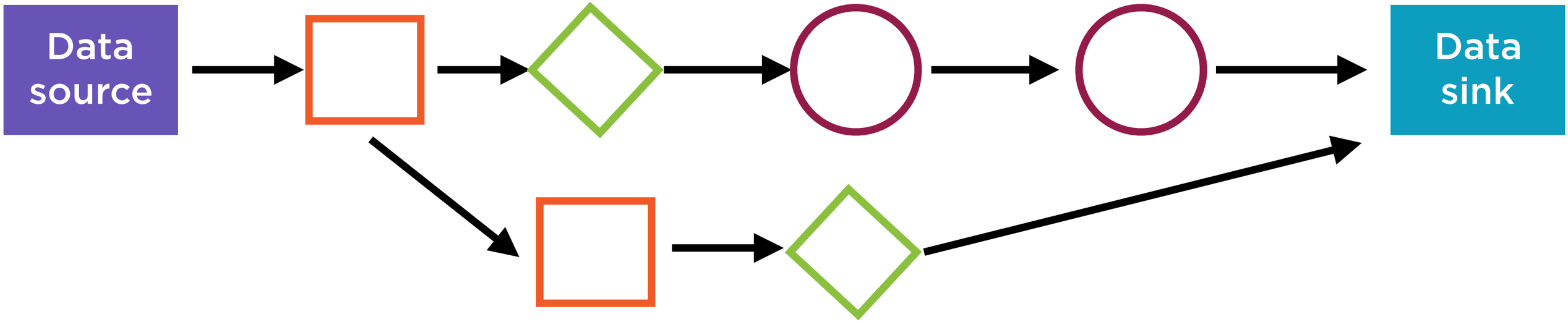
Different transformations applied after data has been split

Demo

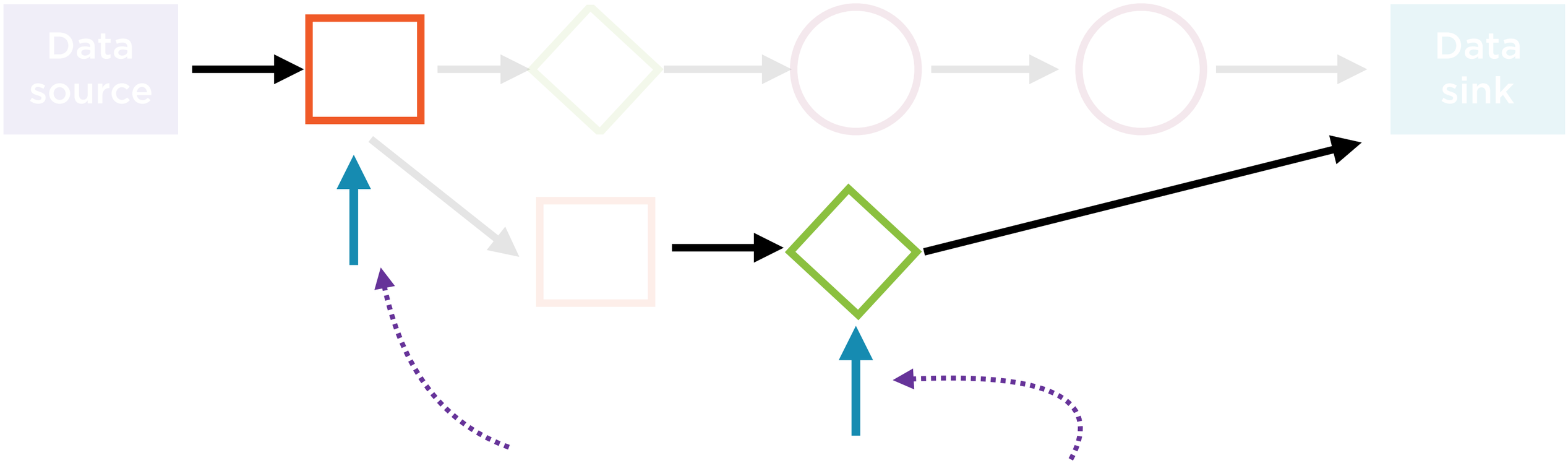
Performing branching operations to split data

Side Inputs

Apache Beam Pipeline

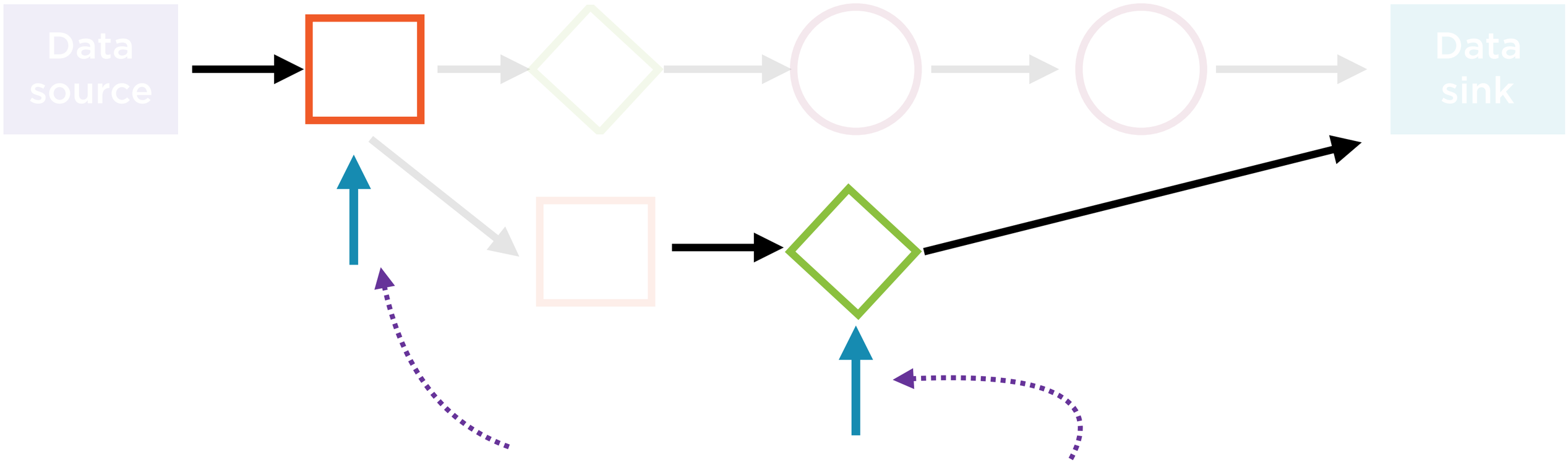


Side Inputs



Inject additional data into some Transform

Side Inputs



Multiple PCollections are processed at that stage

Demo

Build and execute a pipeline using side inputs

Summary

Pipeline as directed acyclic graphs (DAGs)

PCollections as edges, transformations as nodes

Windowing operations

Branching operations

Side inputs into pipelines