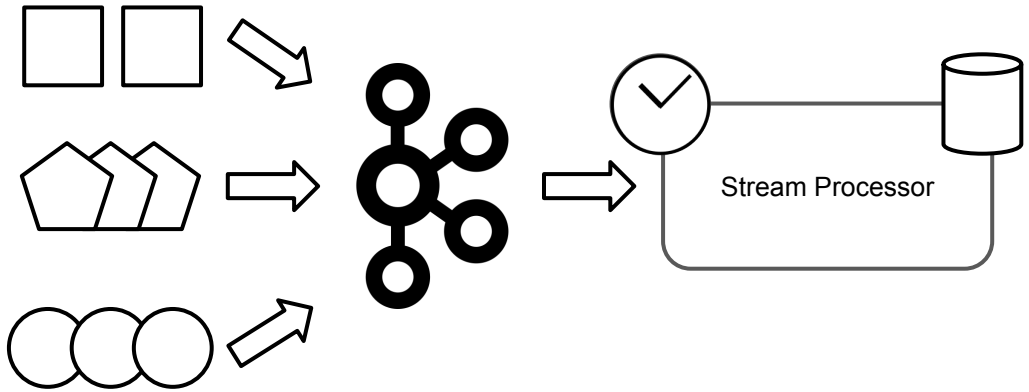
Introduction

Stream Processing Fundamentals

Stream Processing Fundamentals

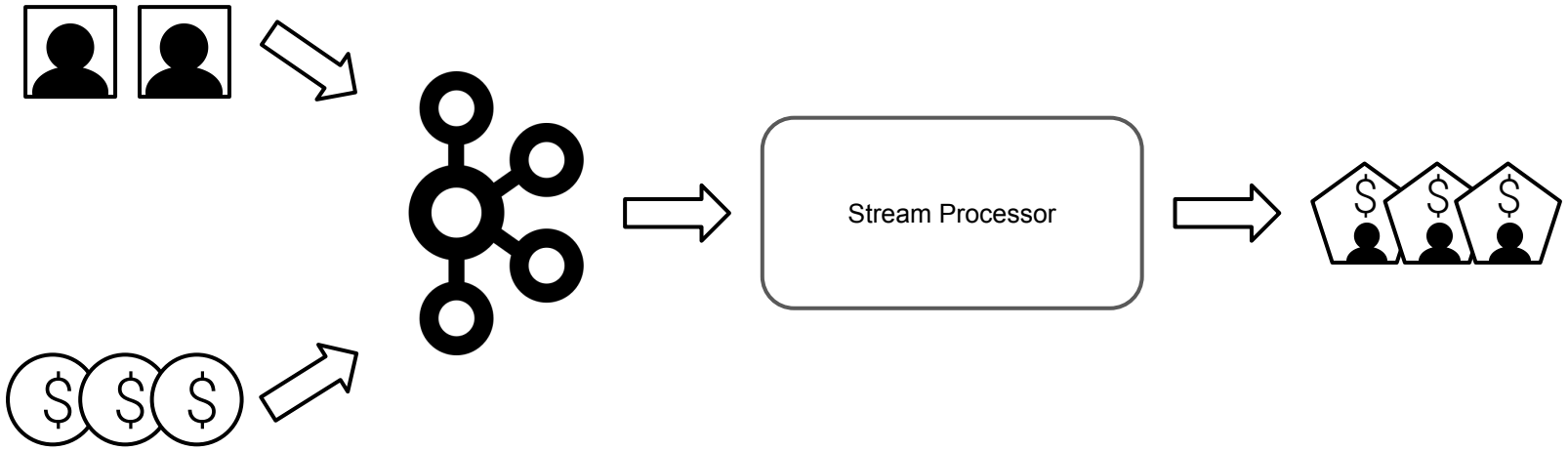
Stream Processing requires developers to consider many factors



Stream Processing Strategies

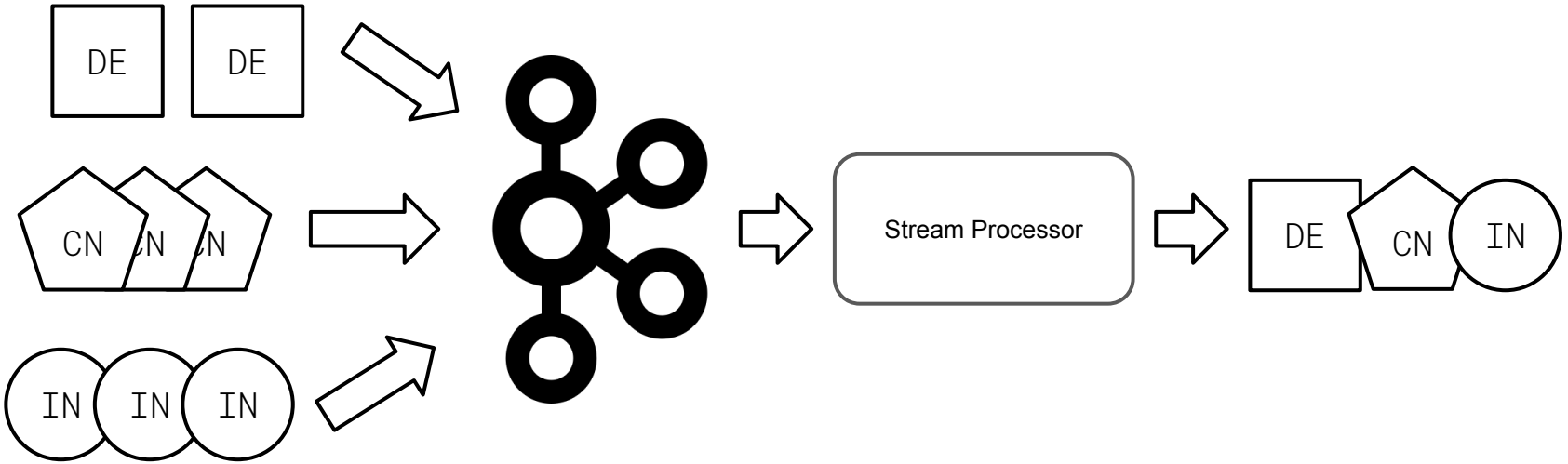
Combining Streams

Combining, or **Joining**, streams is the process of merging their data



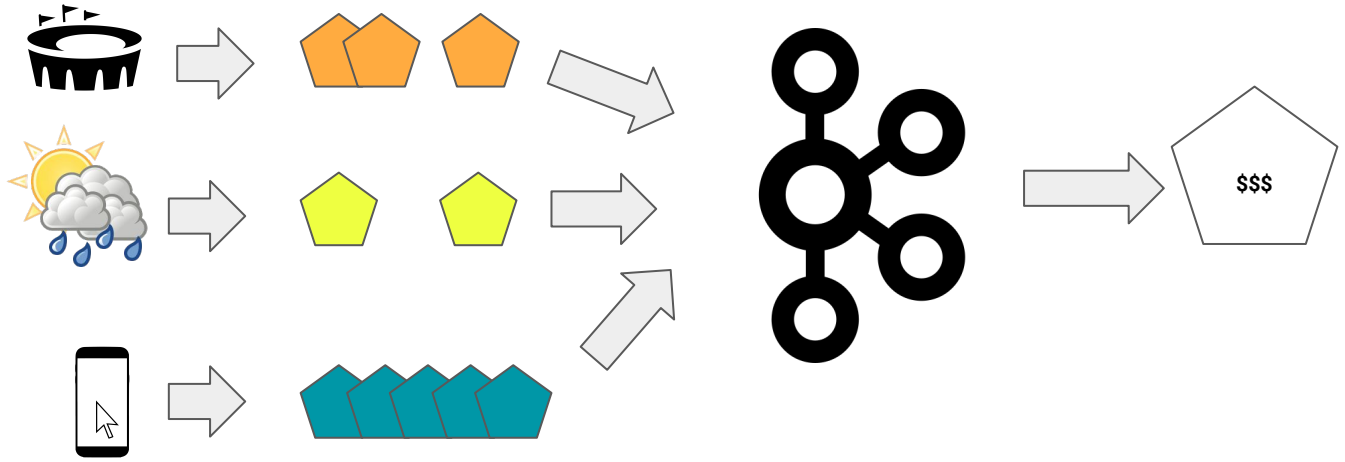
Combining Streams in the Real World

Streams are often joined to aggregate related information



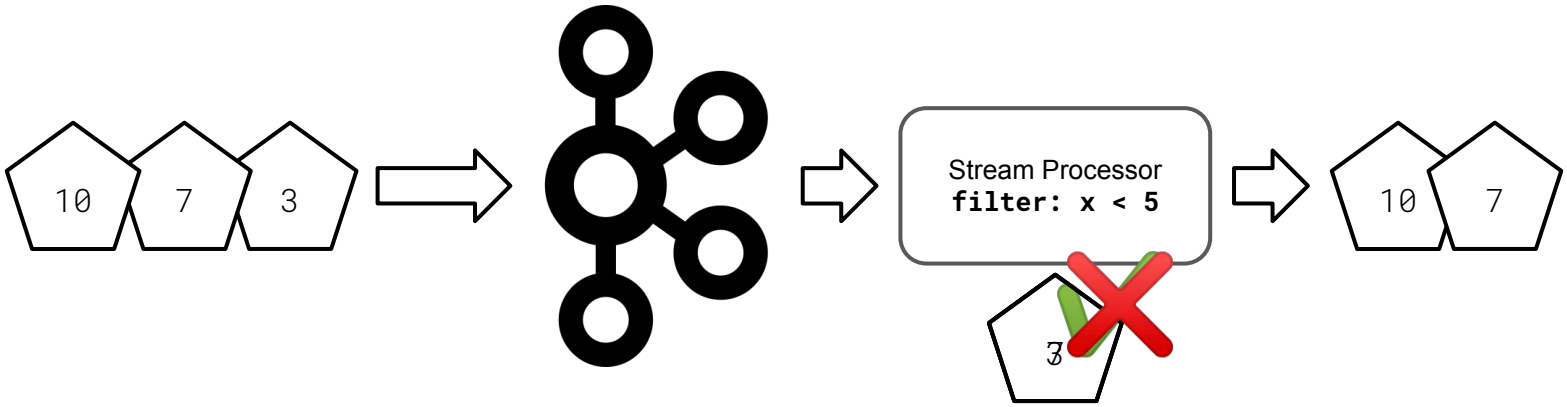
Combining Streams in the Real World

Joins are often paired with aggregations that produce new insights



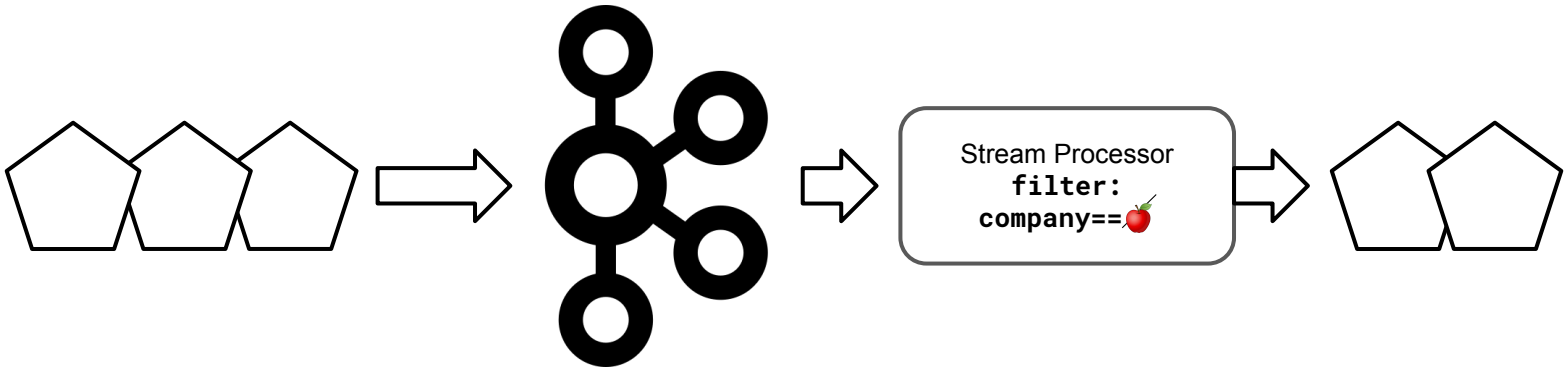
Filtering Streams

The removal of unneeded or unwanted data is **filtering**



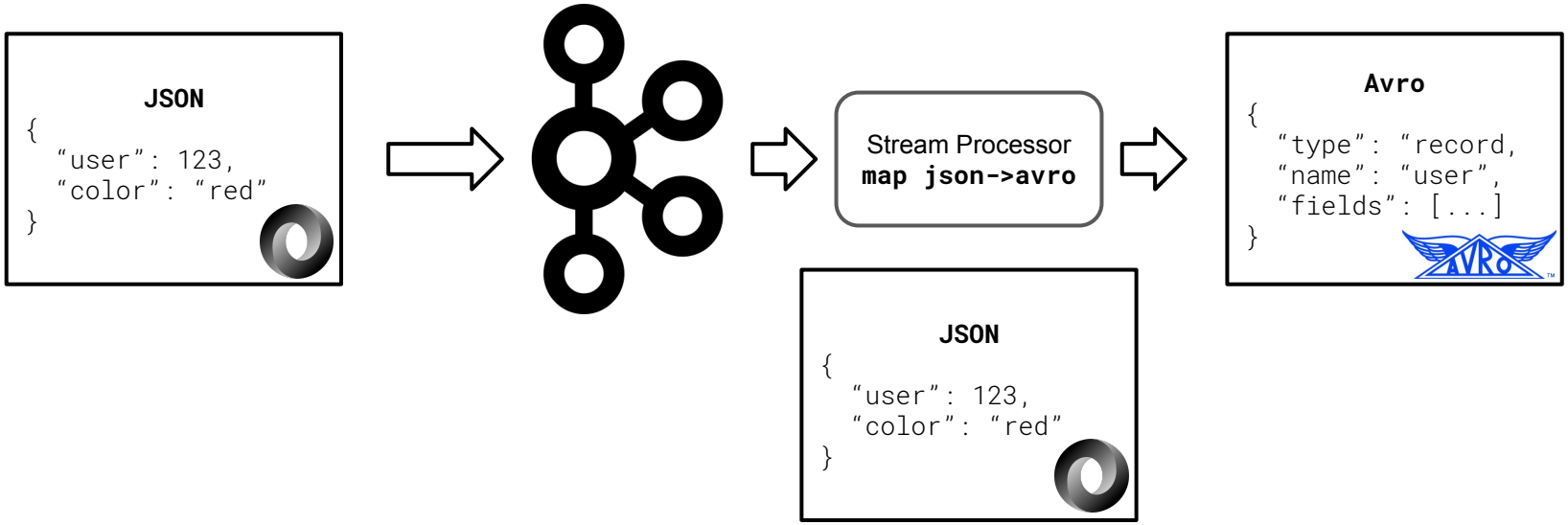
Filtering Streams in the Real World

Filtering is commonly used to create a stream for a more particular use



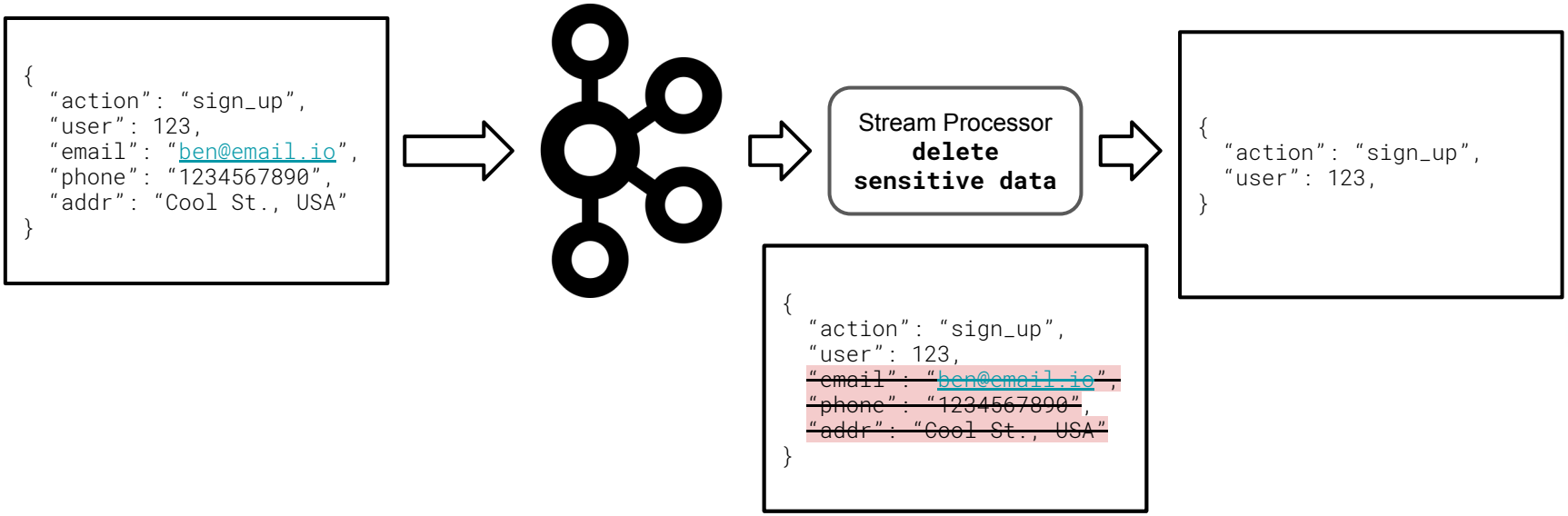
Remapping Streams

Remapping transforms an input event into a different output form



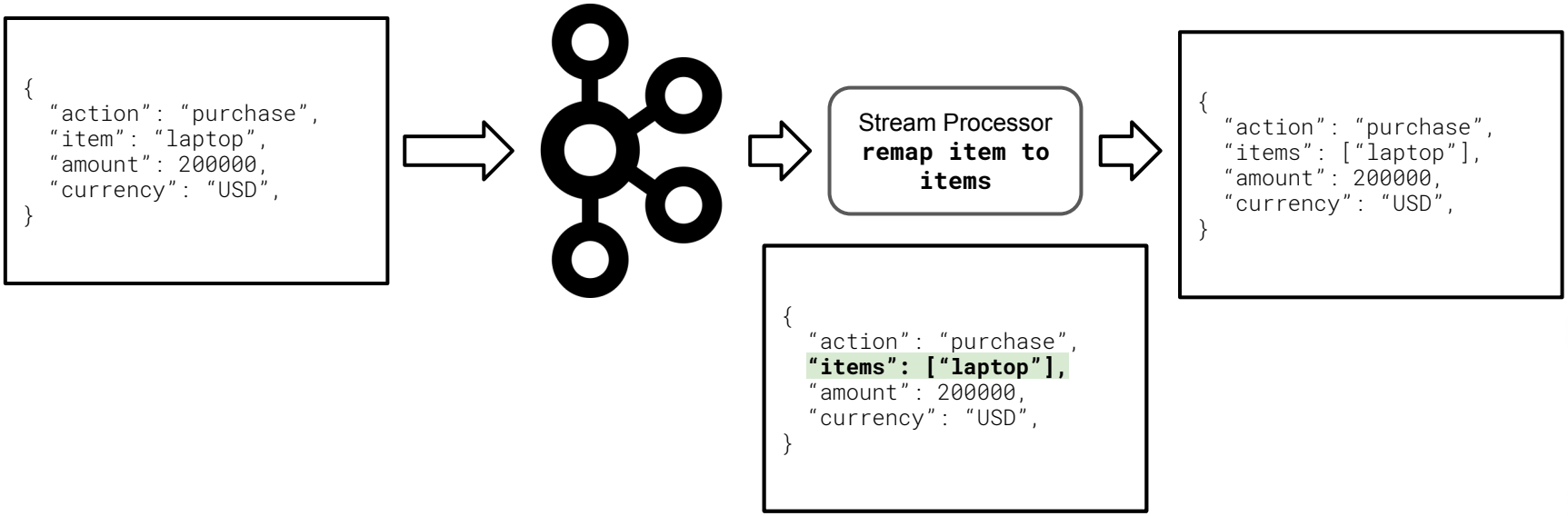
Remapping Streams in the Real World

Remapping is commonly used to remove sensitive data from events



Remapping Streams in the Real World

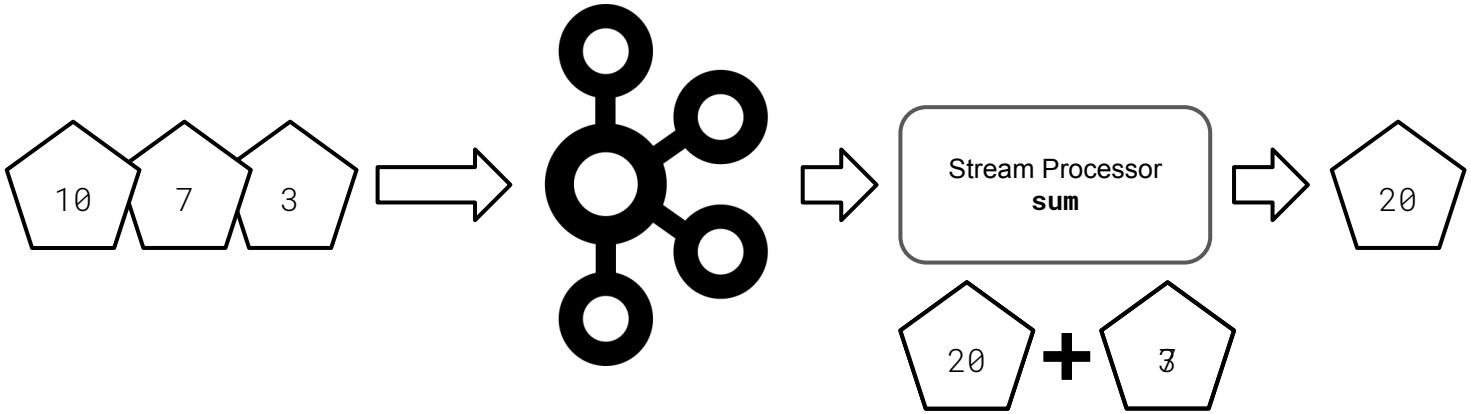
Remapping is commonly used to remove sensitive data from events



Aggregating Streams

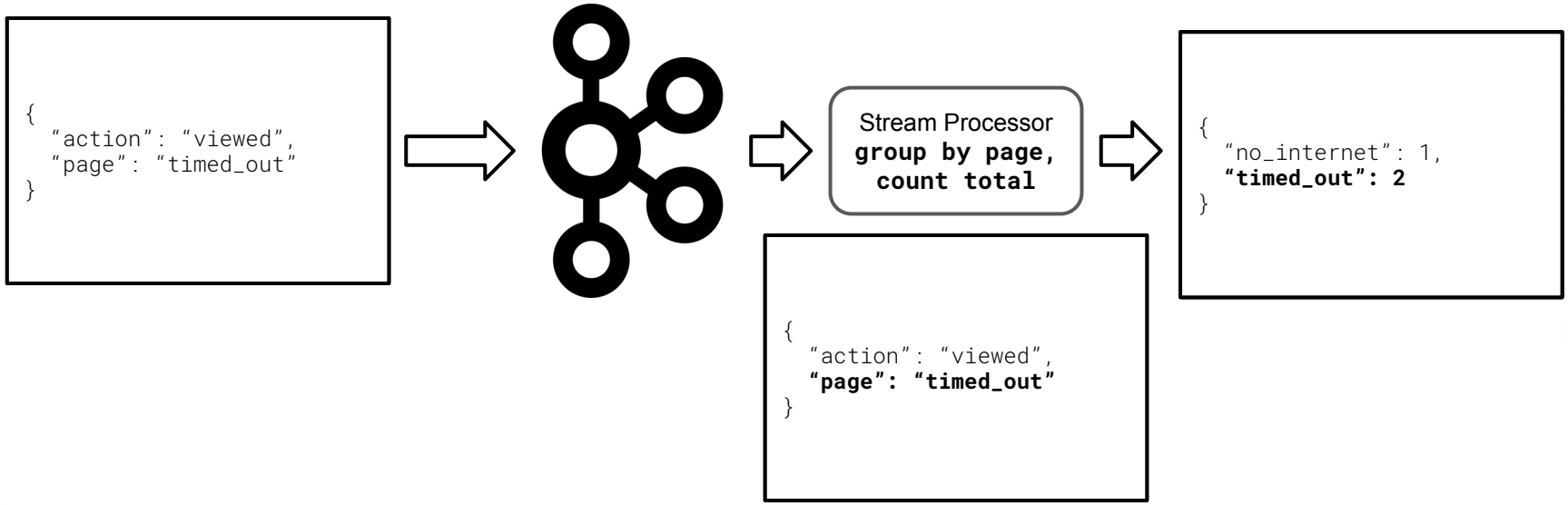
Aggregates take multiple events and emit a new calculated event

Typical aggregate functions are max, min, sum, and histograms



Aggregating Streams in the Real World

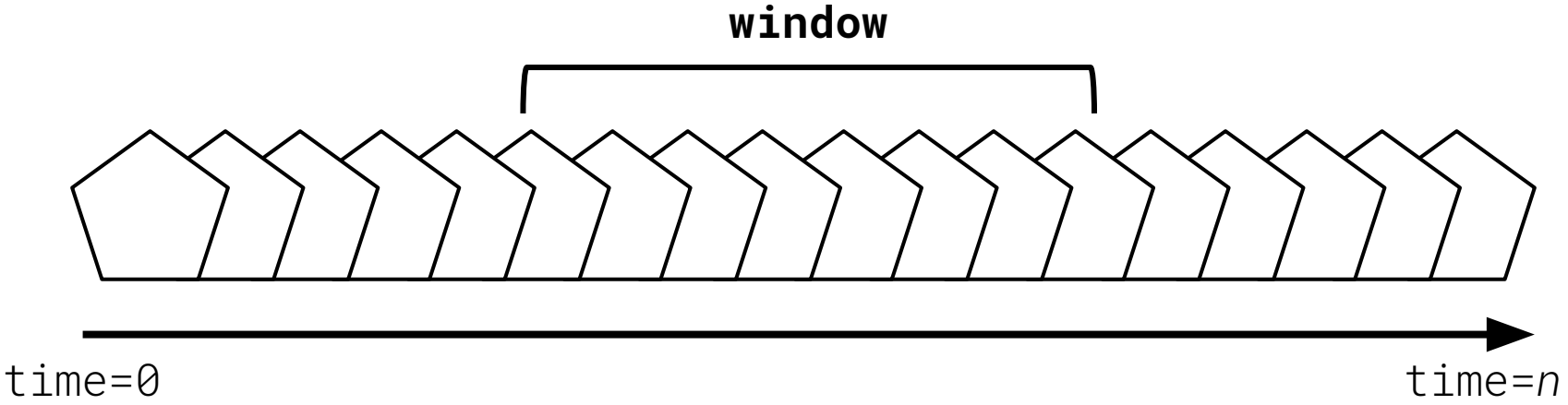
Grouping and counting data by an attribute is a common aggregation



Handling Time

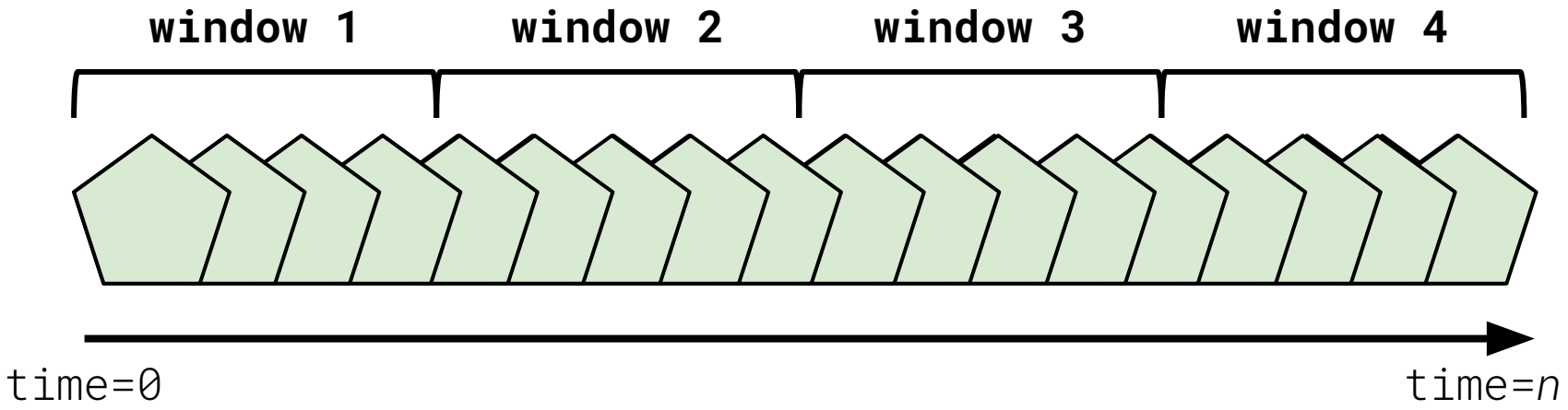
Windowing

A **window** is a period with a **start** and an **end** in which data is gathered for analysis



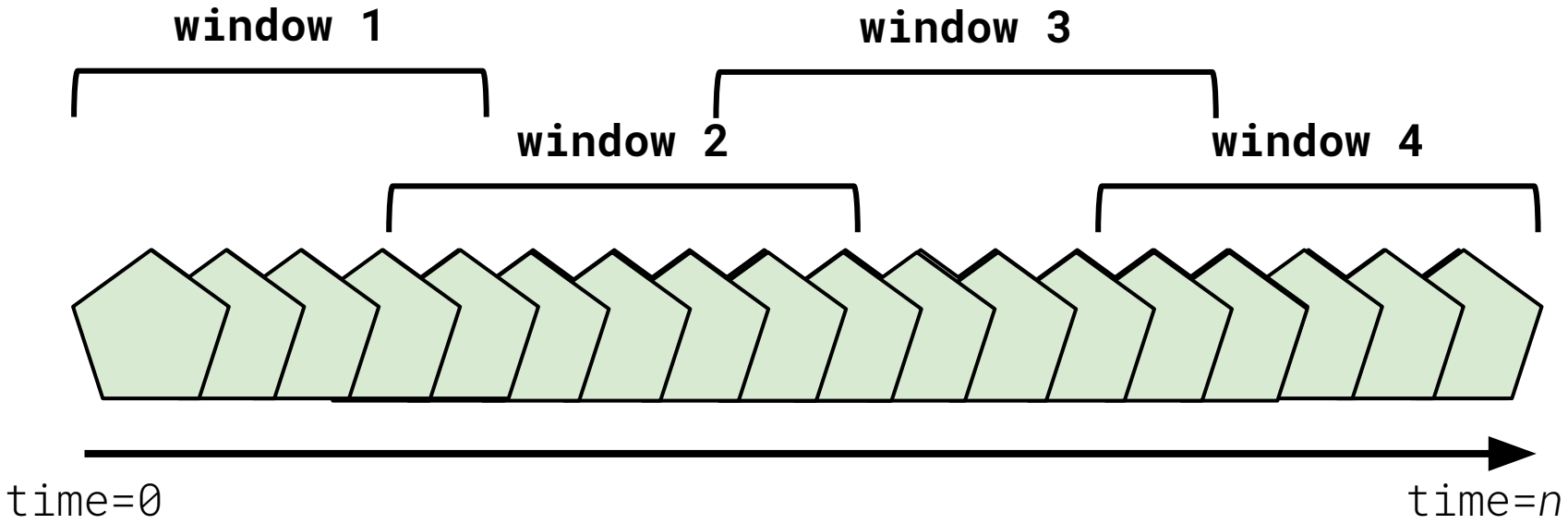
Tumbling Window

A **tumbling window** is a fixed period of time that rolls over after the fixed window has ended



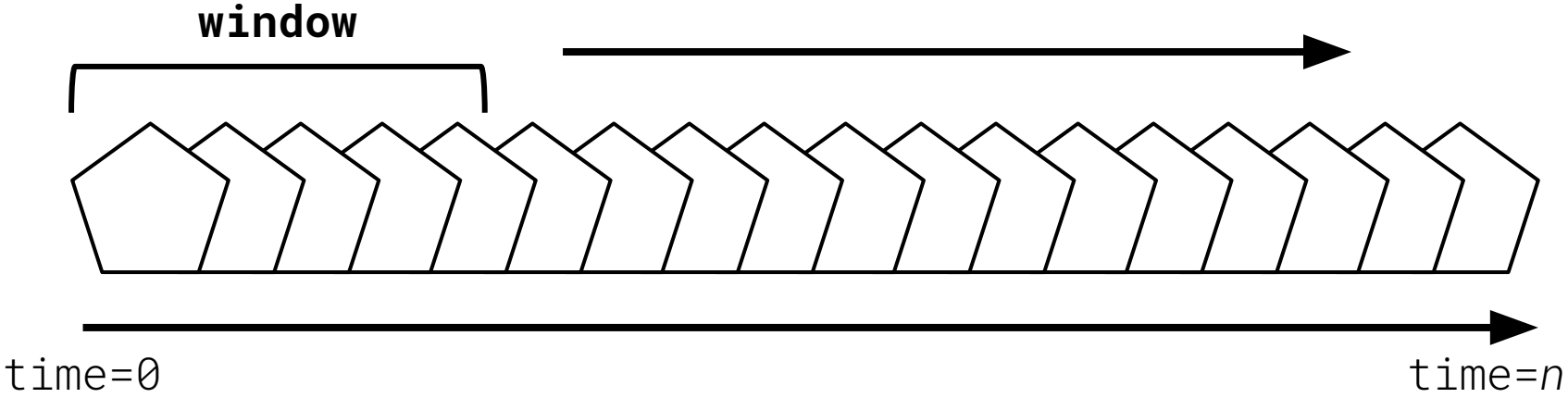
Hopping Window

Hopping windows have a fixed increment which advances the window



Sliding Window

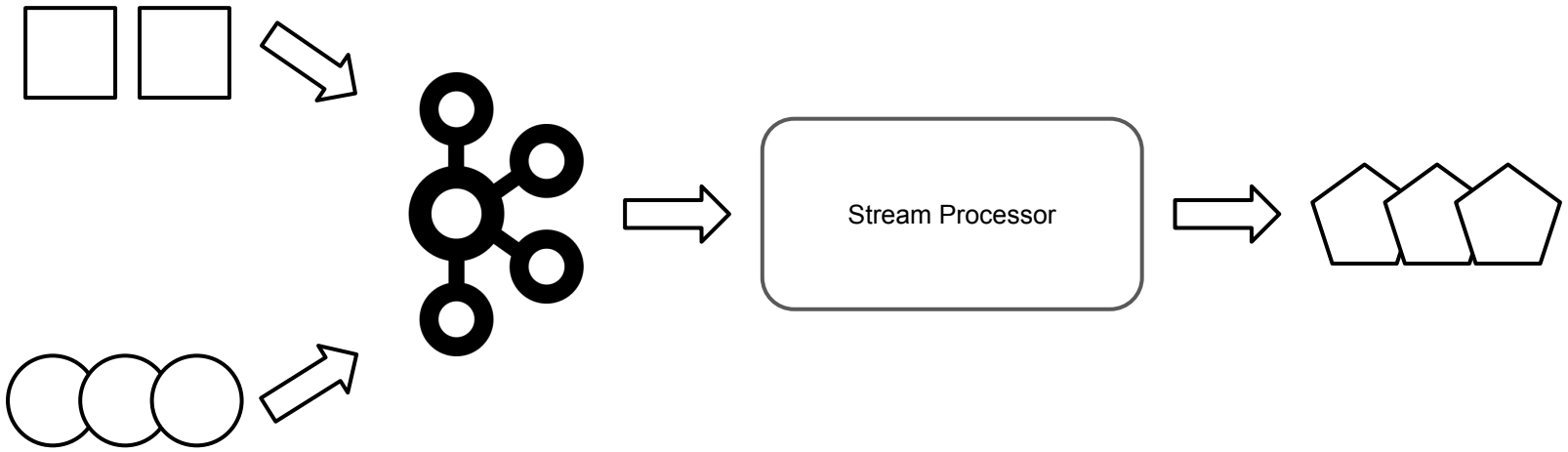
Sliding windows are hopping windows that increment in real-time



Streams and Tables

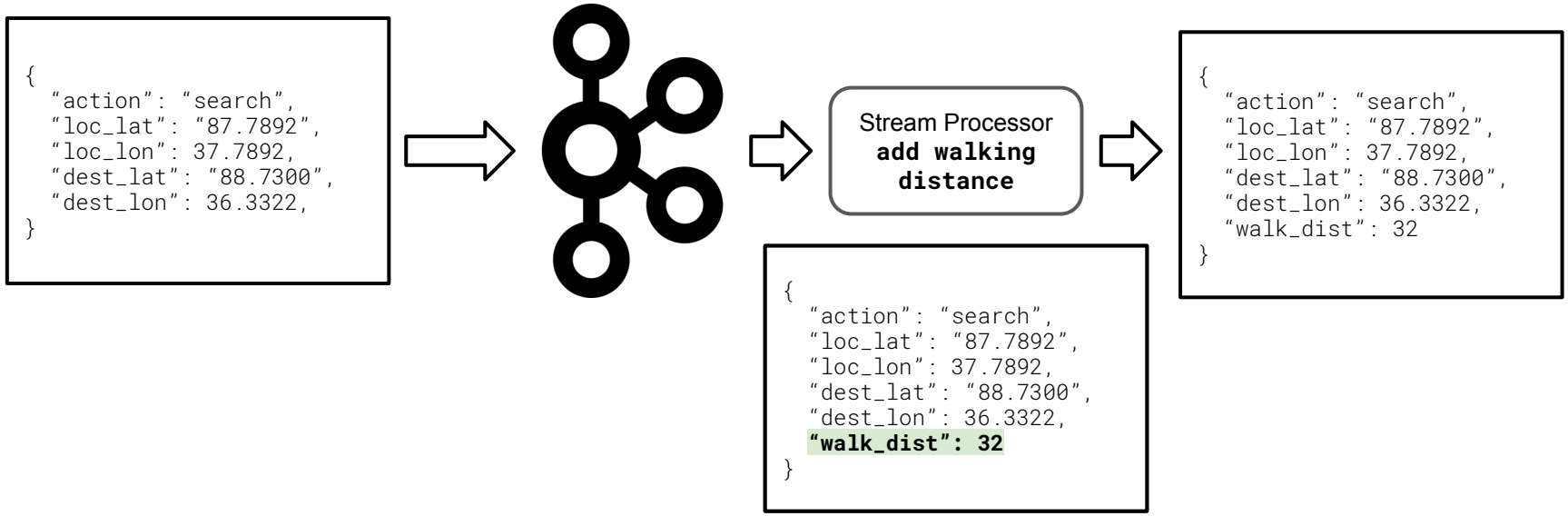
Streams

Streams are an infinite and unbounded sequence of ordered events



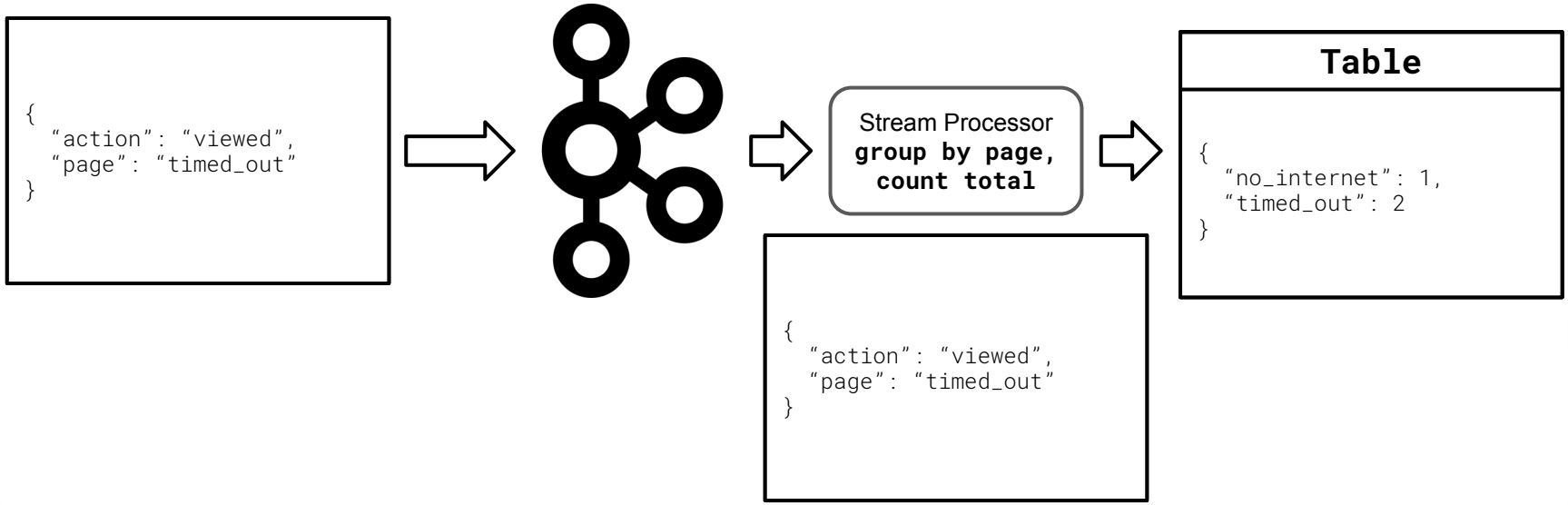
Streams in the Real World

Streams are commonly used to **enrich** data with new fields



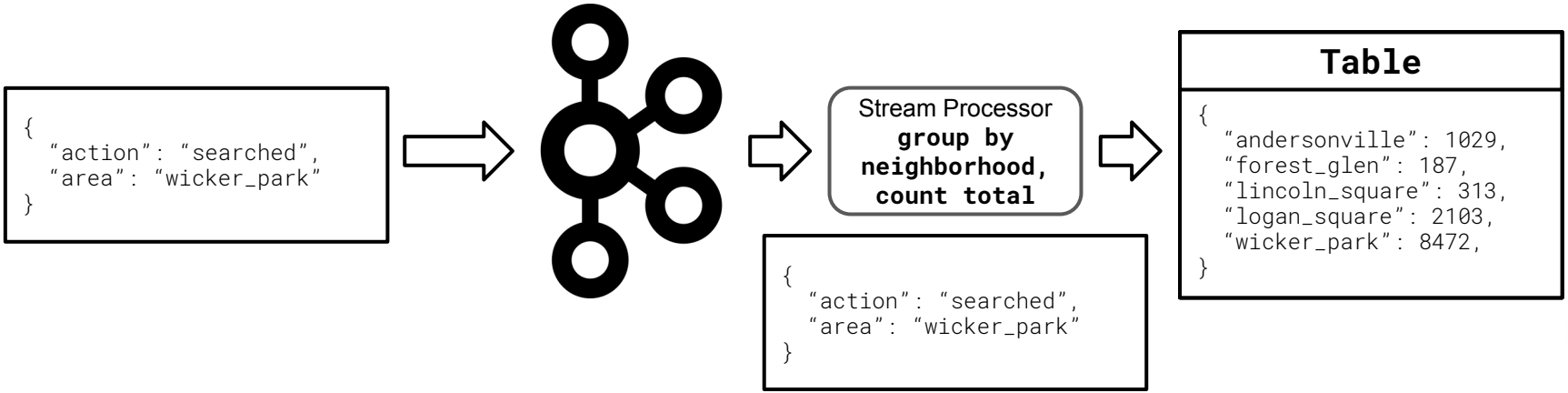
Tables

Streaming tables are the result of stateful aggregations like SUM and COUNT



Tables in the Real World

Tables are frequently used to create running summations of critical data



Streams vs Tables

Complementary Tools

- Streams and tables complement each other
- Streams and tables describe processing *output*
- Streams output an unbounded sequence of events
- Tables output a point-in-time aggregate view

Data Storage

Kafka Changelog

Storage

- All stream processing frameworks require a changelog
- Kafka changelog topic tracks all changes in stream
- Changelog topics are log compacted
- Changelog topic aids in failure tolerance and recovery

RocksDB

Storage

- In-memory storage is default for local state
- In-memory storage is not appropriate for Production
- Always use RocksDB as your local state store
- Always use RocksDB in Production
- RocksDB dramatically speeds reboot/recovery times
- RocksDB is used by all major streaming frameworks

Stream Processing Fundamentals Summary