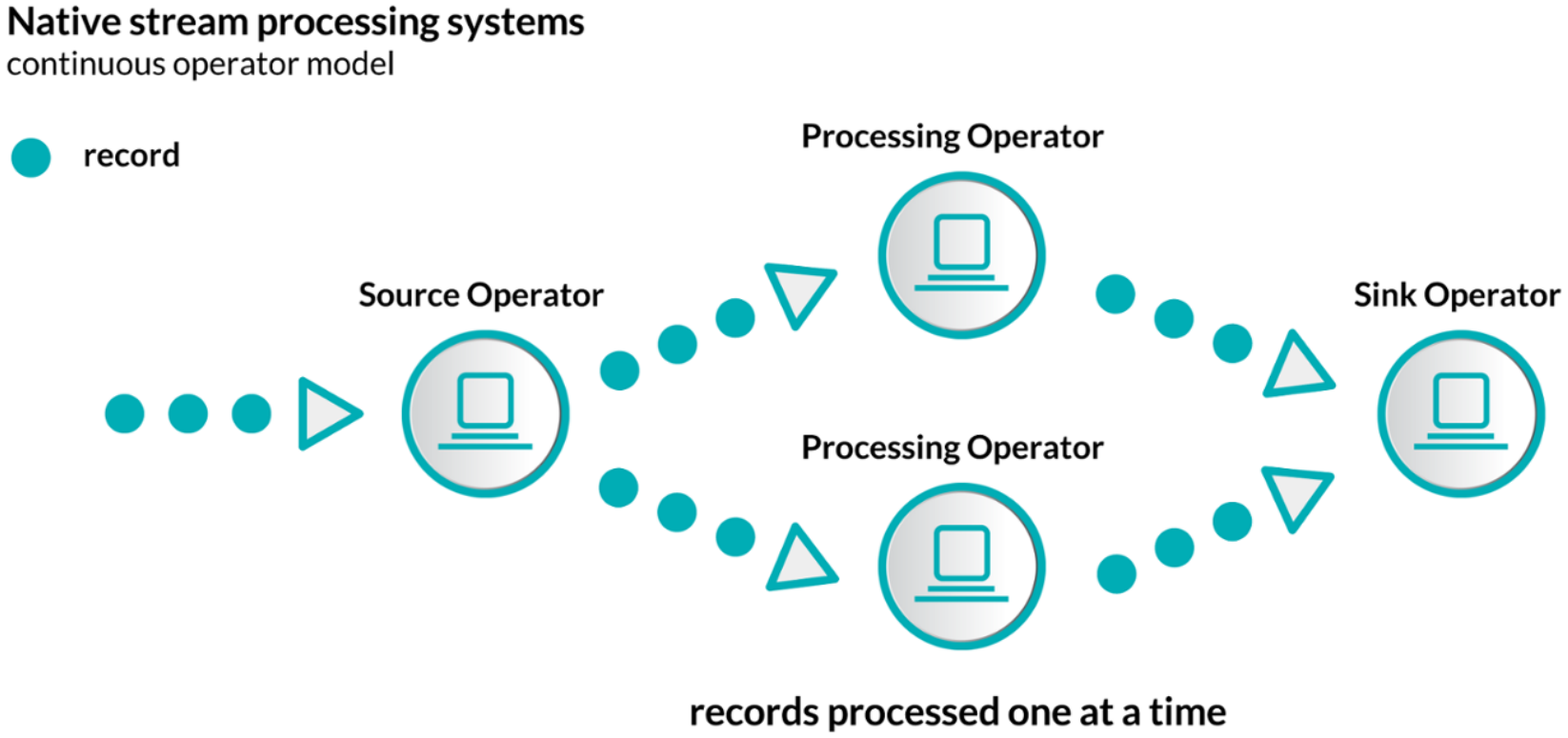
Comparison of streaming processing framework

Distributed processing approach

- Batch Processing
  - has access to all data
  - might compute something big and complex
  - more concerned with throughput that latency
  - higher latencies
- Stream Processing
  - a one-at-a-time processing model
  - data are processed immediately upon arrival
  - computations are relatively simple and generally independent
  - sub-second latency

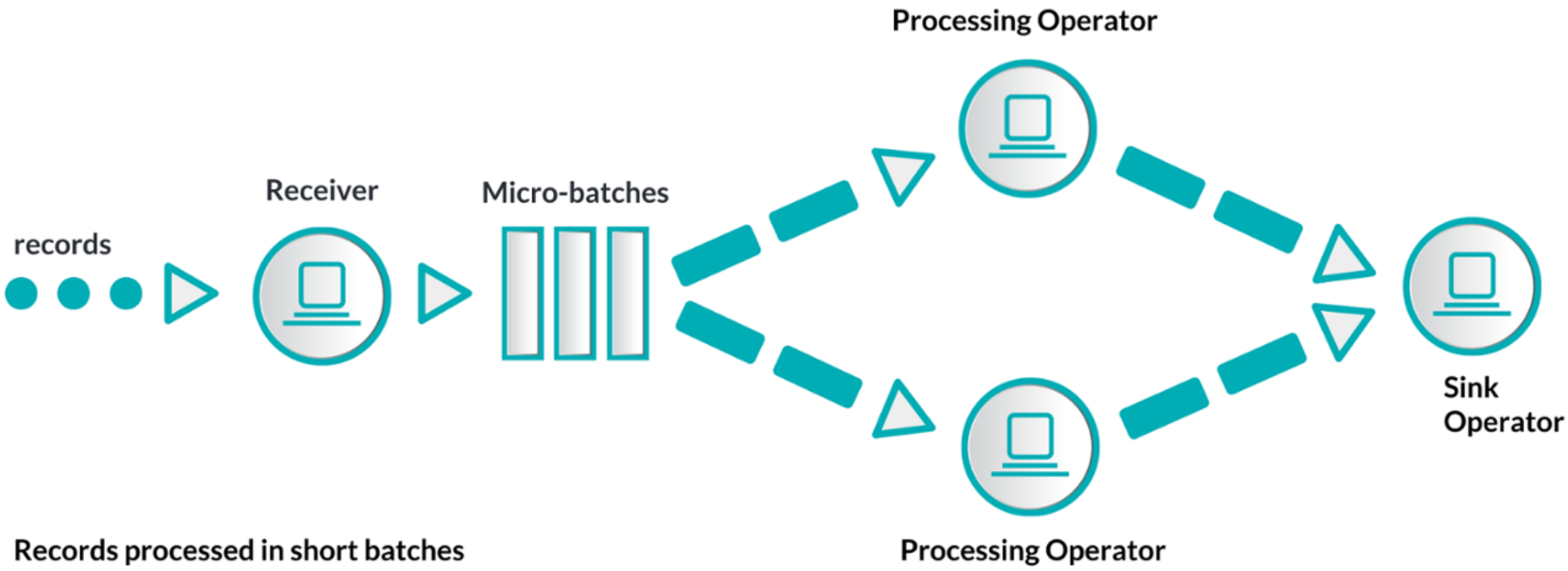
Stream processing approach

- Native stream processing



Stream processing approach

- Micro-batching processing



Stream processing framework

- Storm
- Storm with Trident
- Spark
- Samza
- Flink

Comparing table

	Storm	Trident	Spark	Samza	Flink
Streaming Model	Native	Micro-batching	Micro-batching	Native	Native
API	Compositional		Declarative	Compositional	Declarative
Guarantees	At-least-once	Exactly-once	Exactly-once	At-least-once	Exactly-once
State Management	Not Built-in	Dedicated Operation	Dedicated DStream	Stateful Operation	Stateful Operation
Latency	Very Low	Medium	Medium	Low	Low
Throughput	Low	Medium	High	High	High

API

- Compositional
  - Provide basic building blocks like source or operator
  - Must be tied together to create expected topology
- Declarative
  - System creates and optimize topology itself

API

- Compositional

```
TopologyBuilder builder = new TopologyBuilder();
builder.setSpout("spout", new RandomSentenceSpout(), 5);
builder.setBolt("split", new Split(), 8).shuffleGrouping("spout");
builder.setBolt("count", new WordCount(), 12).fieldsGrouping("split", new Fields("word"));

```

- Declarative

```
val conf = new SparkConf().setAppName("wordcount")
val ssc = new StreamingContext(conf, Seconds(1))

val text = ...

val counts = text.flatMap(line => line.split(" "))
 .map(word => (word, 1))
 .reduceByKey(_ + _)

counts.print()

ssc.start()
ssc.awaitTermination()
```

Guarantee

- Message Delivery Guarantees
  - At most once: data may be lost
  - At least once: data may be duplicated
  - Exactly once: data neither lost nor duplicated

Comparing table

	Storm	Trident	Spark	Samza	Flink
Streaming Model	Native	Micro-batching	Micro-batching	Native	Native
API	Compositional		Declarative	Compositional	Declarative
Guarantees	At-least-once	Exactly-once	Exactly-once	At-least-once	Exactly-once
State Management	Not Built-in	Dedicated Operation	Dedicated DStream	Stateful Operation	Stateful Operation
Latency	Very Low	Medium	Medium	Low	Low
Throughput	Low	Medium	High	High	High