Processing Data with the Streaming API

Justin Pihony

@JustinPihony|justin-pihony.blogspot.com

Course Overview

DataFrames

Datasets

Spark Streaming

Optimizing Towards Fast Data

Module Overview

Spark Streaming

- Landscape
- Understanding
- Doing
- Working with State
- Monitoring

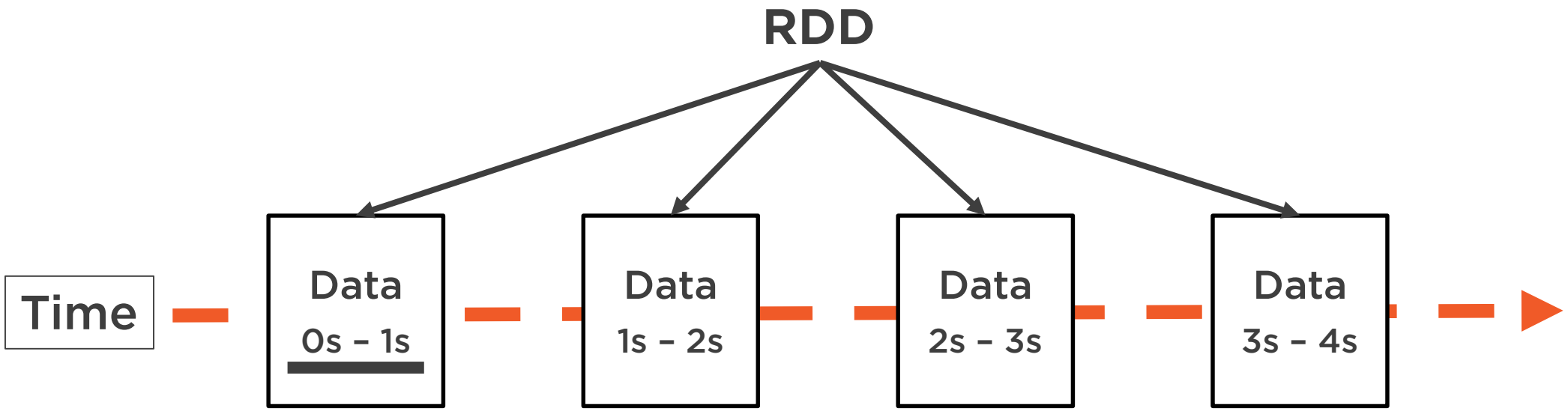
Streaming Landscape Review

Streaming Landscape

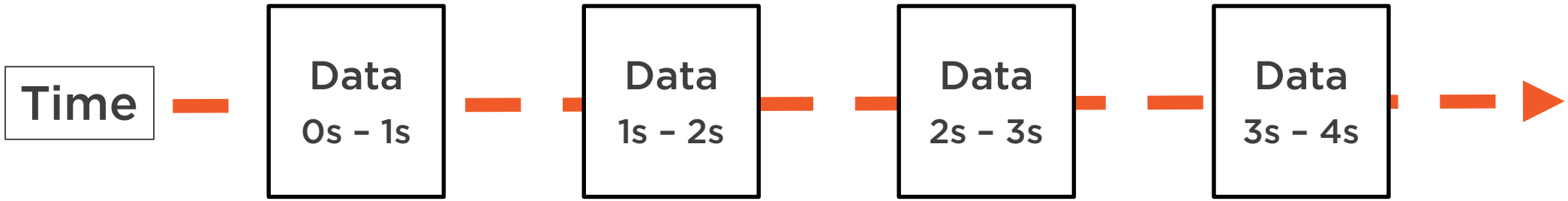
Streaming

Understanding the Concepts

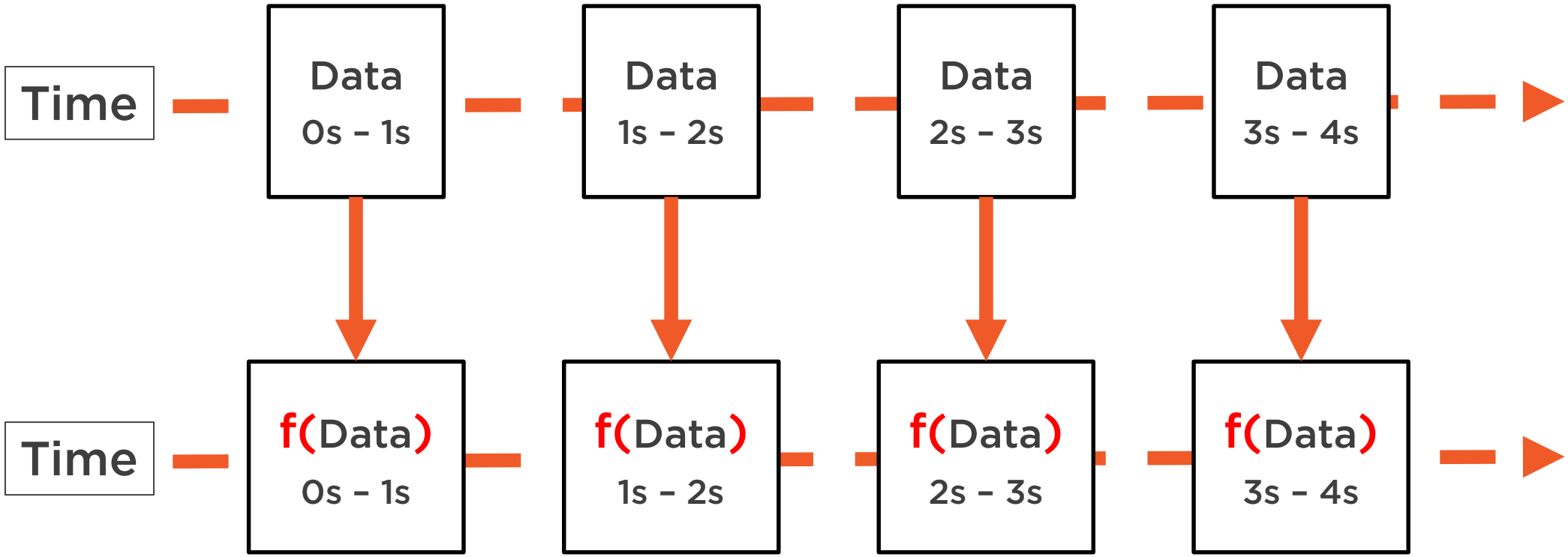
DStreams



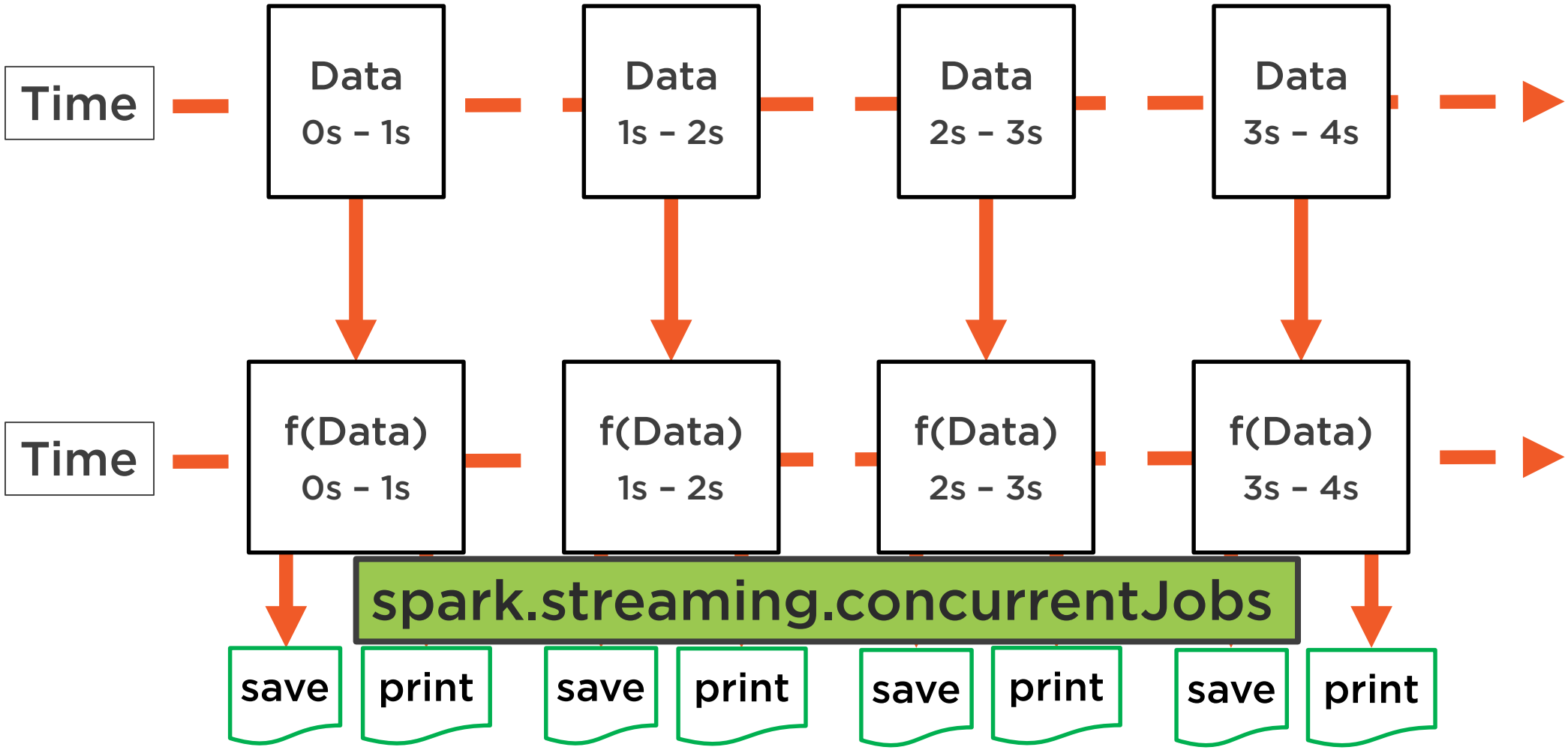
DStreams



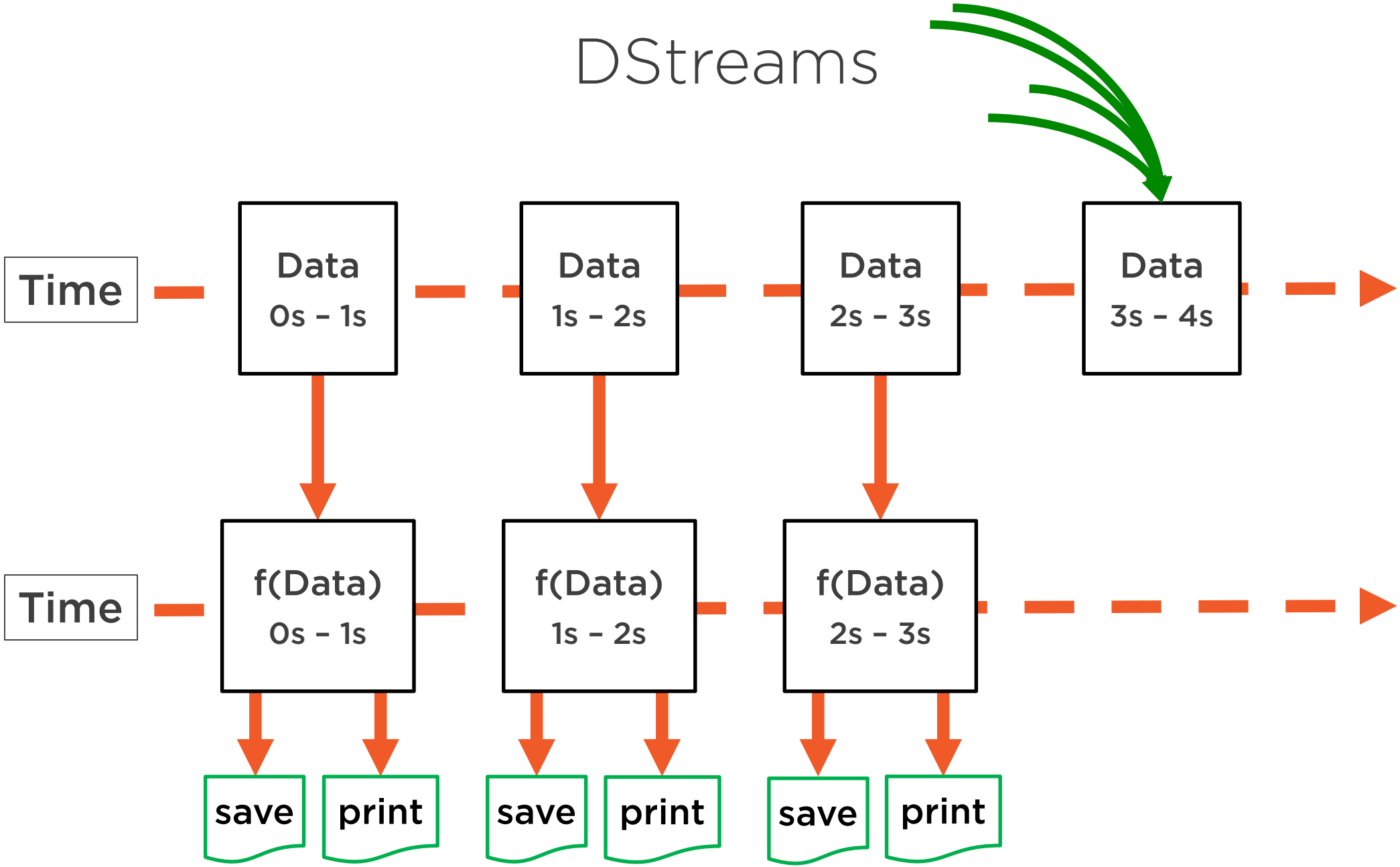
DStreams


Time
Data
0s – 1s
Data
1s – 2s
Data
2s – 3s
Data
3s – 4s
Time
f(Data)
0s – 1s
f(Data)
1s – 2s
f(Data)
2s – 3s
f(Data)
3s – 4s

DStreams


Time
Data
0s – 1s
Data
1s – 2s
Data
2s – 3s
Data
3s – 4s
Time
f(Data)
0s – 1s
f(Data)
1s – 2s
f(Data)
2s – 3s
f(Data)
3s – 4s
spark.streaming.concurrentJobs
save
print
save
print
save
print
save
print

DStreams


Time
Data
0s – 1s
Data
1s – 2s
Data
2s – 3s
Data
3s – 4s
Time
f(Data)
0s – 1s
f(Data)
1s – 2s
f(Data)
2s – 3s
save
print
save
print
save
print

DStream “RDD” API

RDD-like DStream Methods

- map(Partitions) & flatMap
- filter
- reduce
- glom
- repartition
- context
- cache/persist StorageLevel.MEMORY_ONLY_SER

RDD-like DStream Methods

- map(Partitions) & flatMap
- filter
- reduce
- glom
- repartition
- context
- cache/persist StorageLevel.MEMORY_ONLY_SER_2

RDD-like DStream Methods

- `map(Partitions) & flatMap`
- `filter`
- `reduce`
- `glom`
- `repartition`
- `context`
- `cache/persist` **spark.cleaner.ttl** **(ttl=time to live)**

RDD-like DStream Methods

- map(Partitions) & flatMap
- filter
- reduce
- glom
- repartition
- context
- cache/persist
- (ssc.)union
- ssc.transform/transformWith
- saveAs
  - ObjectFiles
  - TextFiles
  - (NewAPI)HadoopFiles

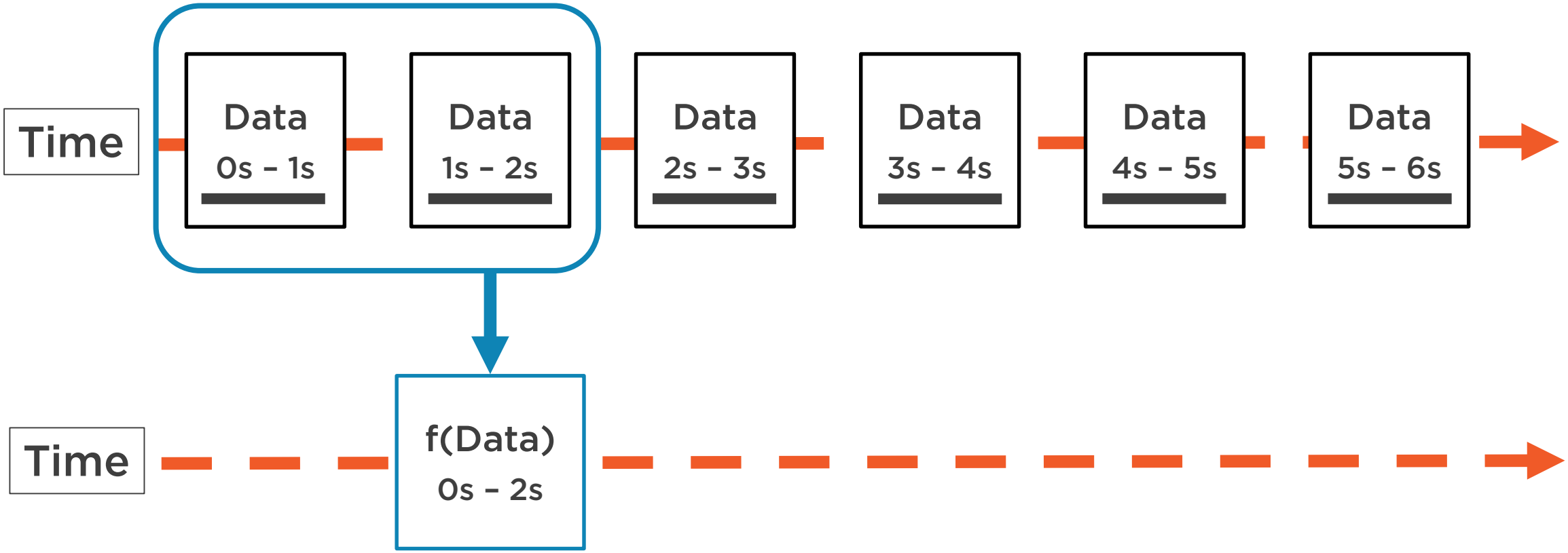
PairDStream

- mapValues & flatMapValues
- groupByKey
- reduceByKey
- combineByKey
- cogroup
- join
  - fullOuterJoin
  - leftOuterJoin
  - rightOuterJoin

Stateful Streaming

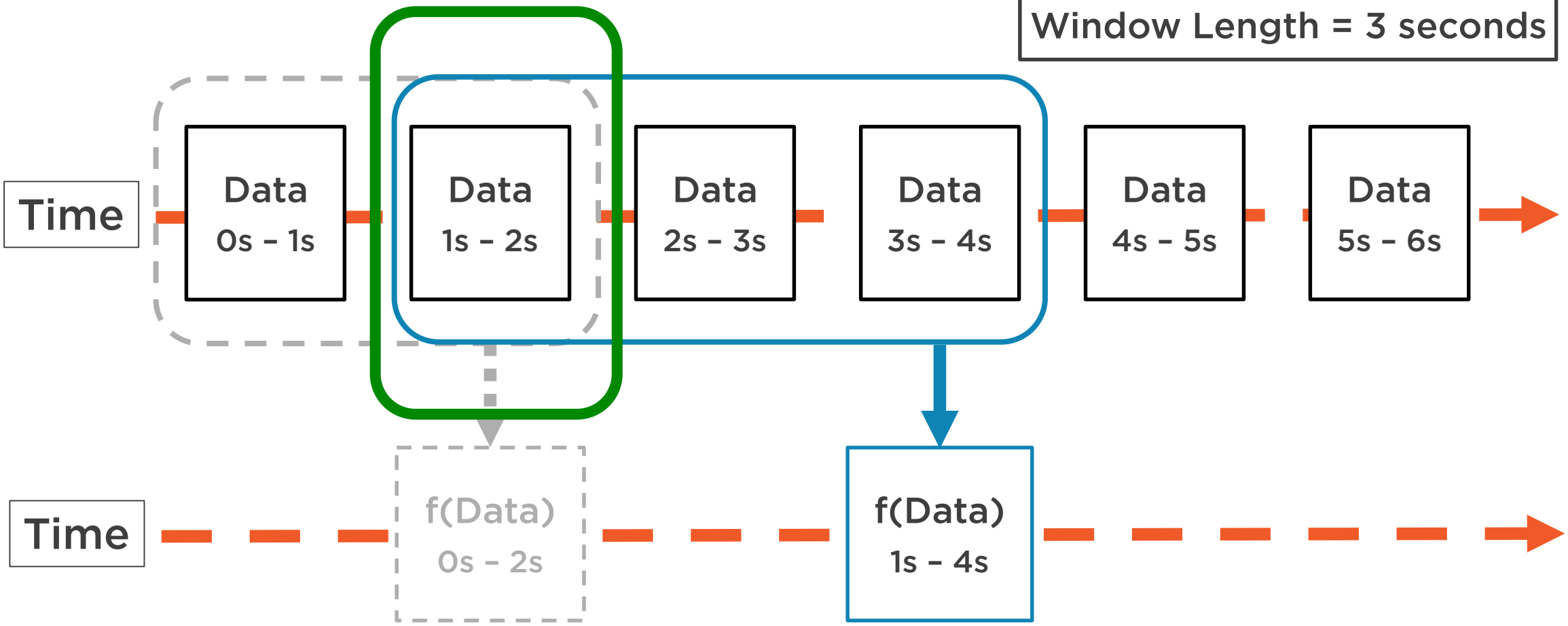
Windows

Batch Size = 1 seconds
Slide Interval = 2 seconds



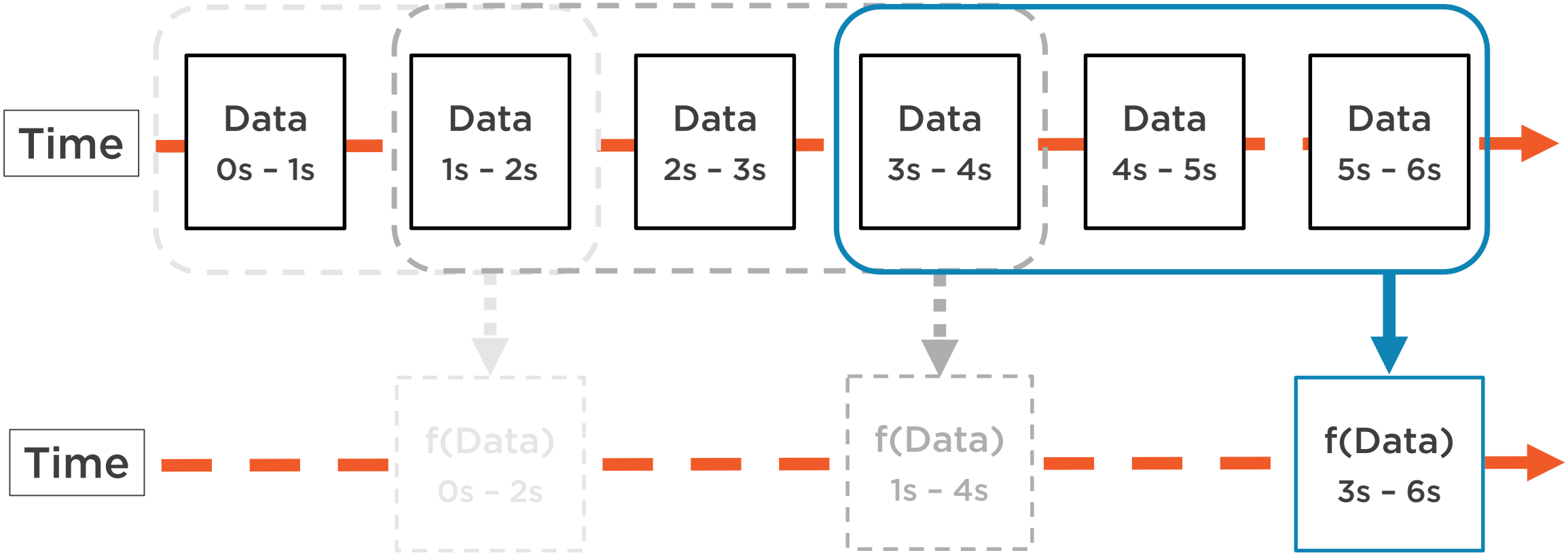
Windows

Batch Size = 1 seconds
Slide Interval = 2 seconds
Window Length = 3 seconds



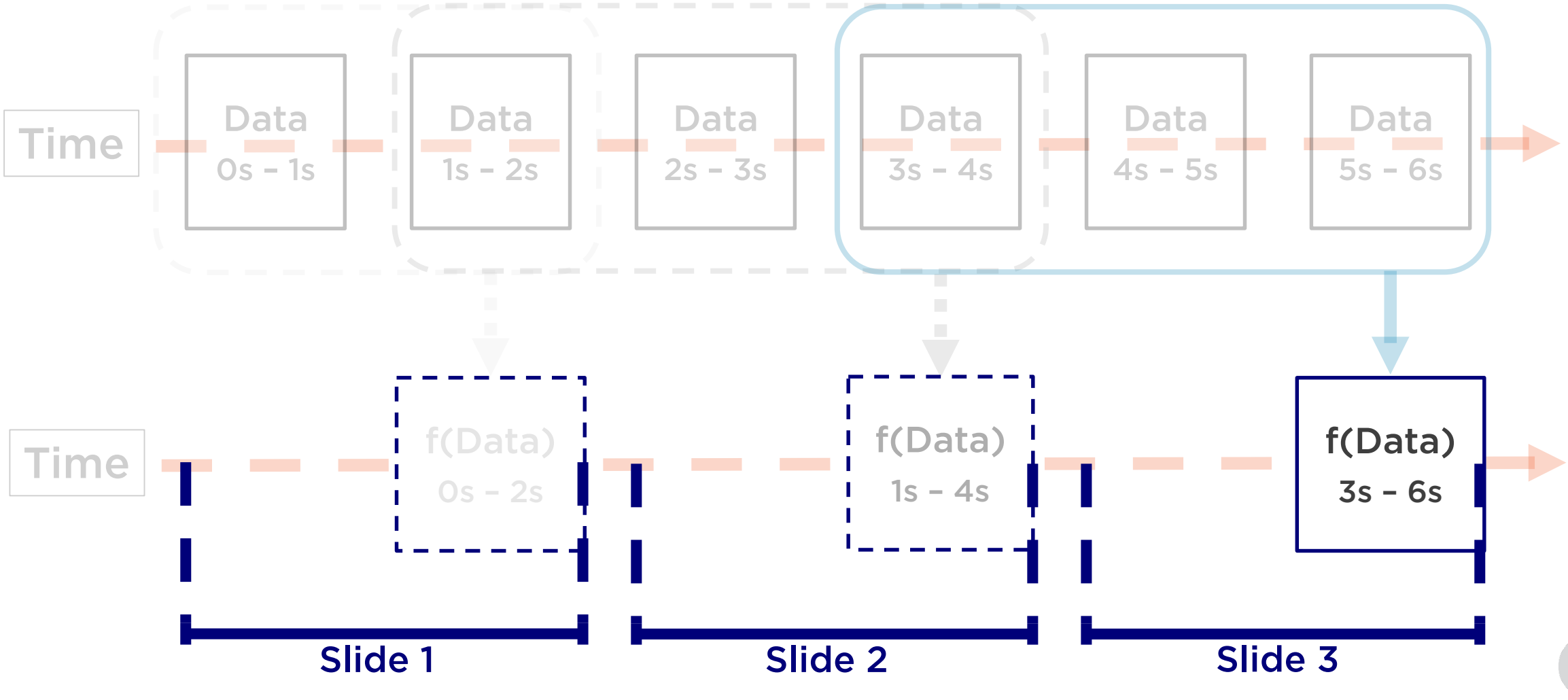
Windows

Batch Size = 1 seconds
Slide Interval = 2 seconds
Window Length = 3 seconds



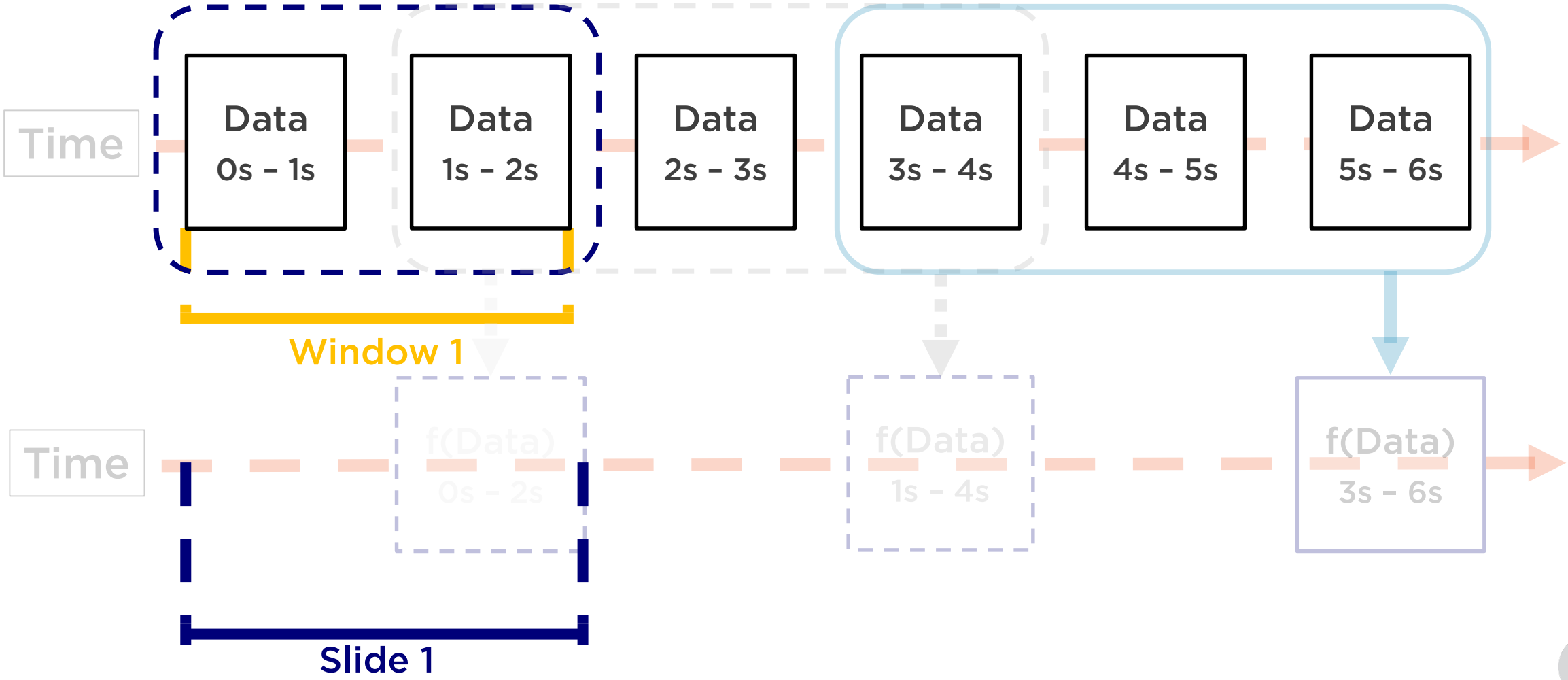
Windows

Batch Size = 1 seconds
Slide Interval = 2 seconds
Window Length = 3 seconds



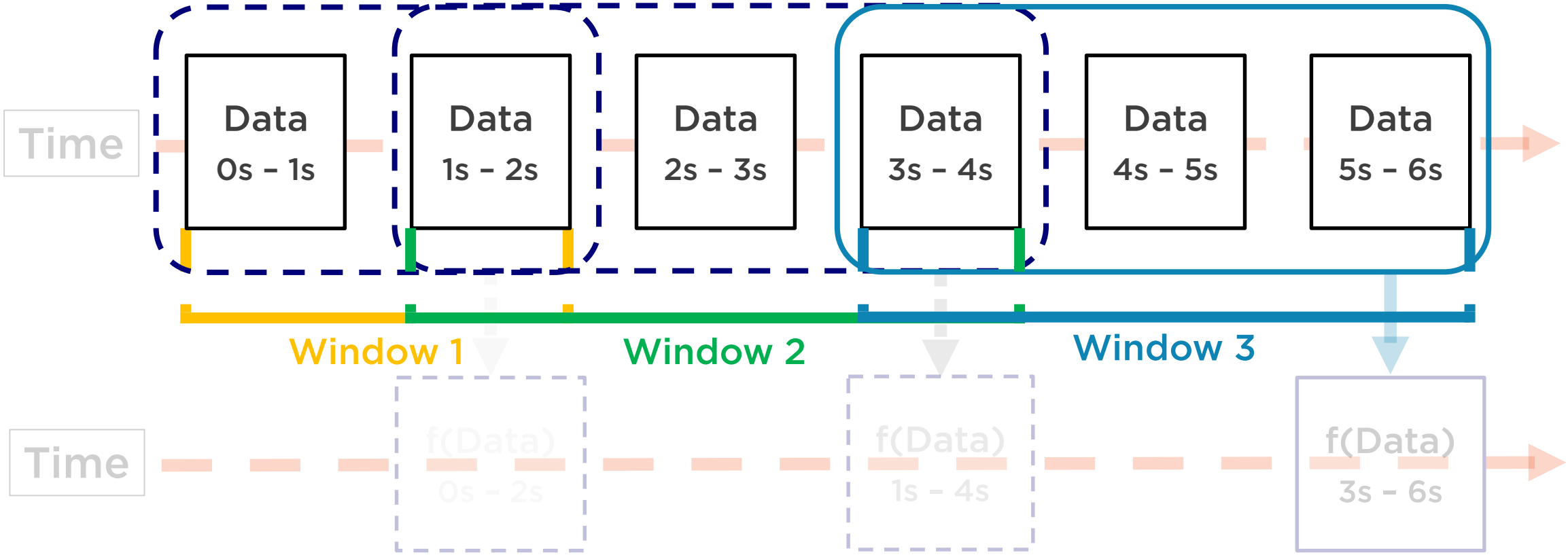
Windows

Batch Size = 1 seconds
Slide Interval = 2 seconds
Window Length = 3 seconds

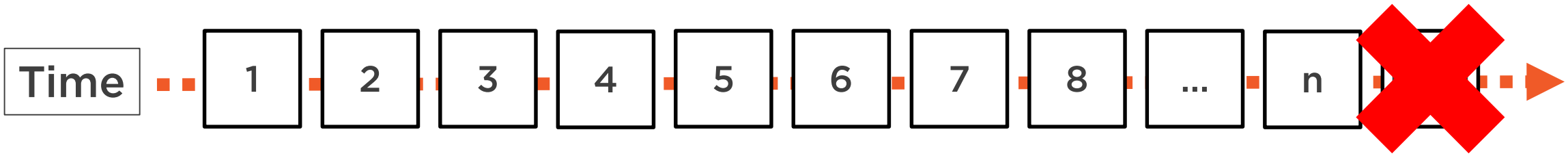


Windows

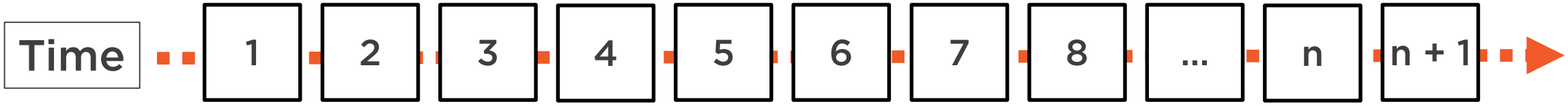
Batch Size = 1 seconds
Slide Interval = 2 seconds
Window Length = 3 seconds



Checkpointing



Checkpointing



Checkpointing

Checkpointing

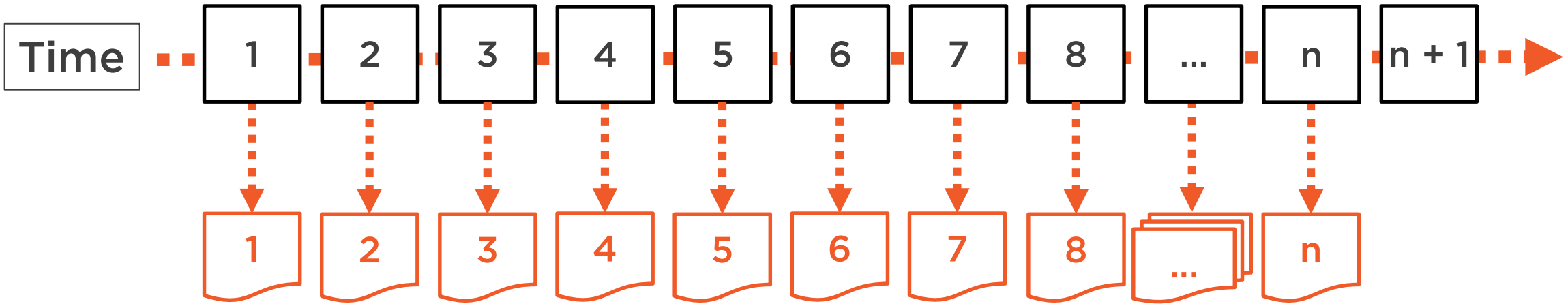
Checkpointing

Checkpointing

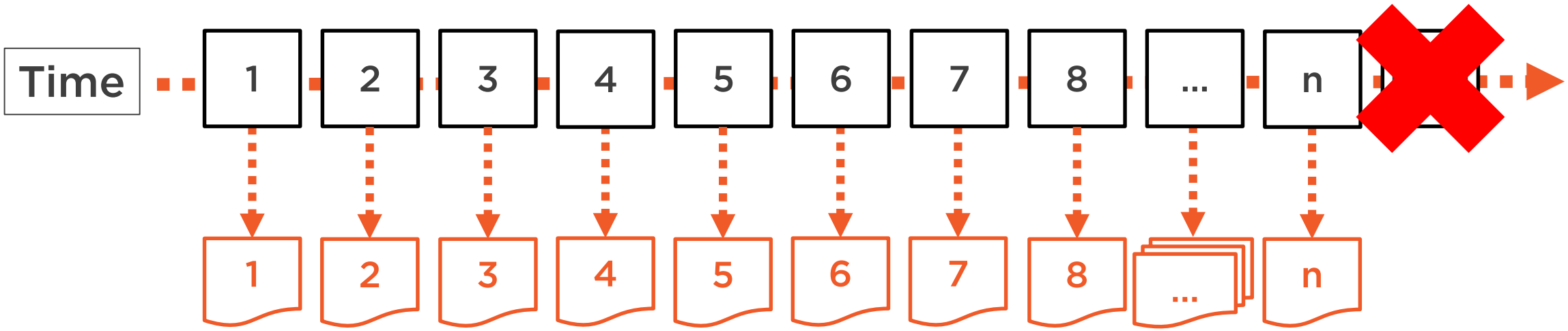
Checkpointing

Checkpointing

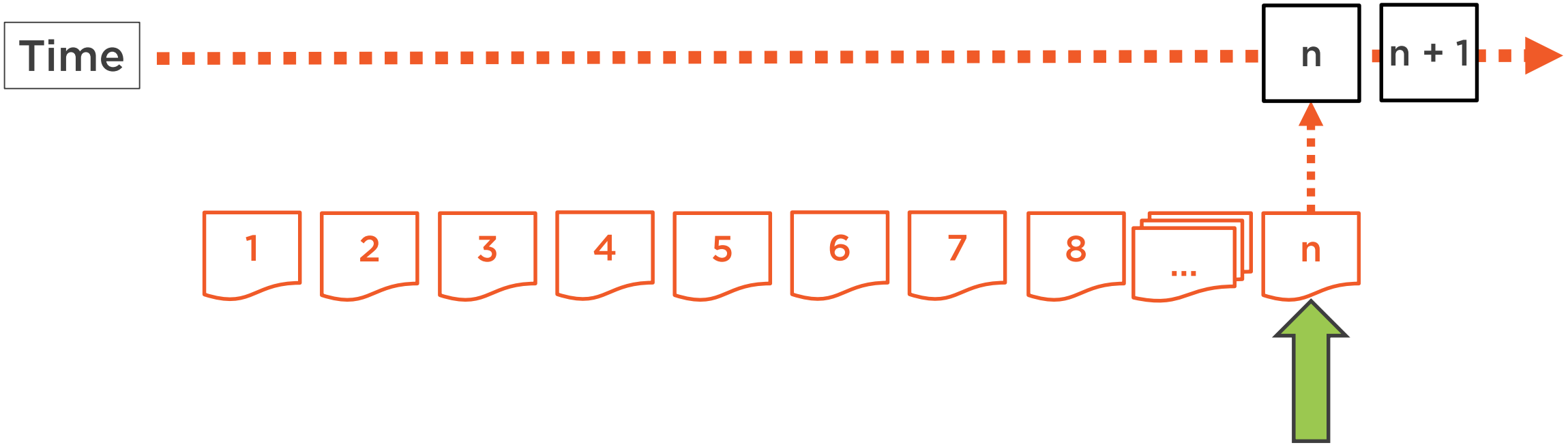
Checkpointing



Checkpointing



Checkpointing



Resources

- Comparison of Apache Stream Processing Frameworks: Petr Zapletal
  - cakesolutions.net/teamblogs/comparison-of-apache-stream-processing-frameworks-part-1
- Diving into Apache Spark Streaming's Execution Model: Databricks
  - databricks.com/blog/2015/07/30/diving-into-apache-spark-streamings-execution-model
- Deep Dive with Spark Streaming: Tathagata Das
  - youtube.com/watch?v=D1knCQZQQnw
- New Visualizations for Understanding Apache Spark Streaming Applications: Databricks
  - databricks.com/blog/2015/07/08/new-visualizations-for-understanding-apache-spark-streaming-applications
- Faster Stateful Stream Processing in Apache Spark Streaming: Databricks
  - databricks.com/blog/2016/02/01/faster-stateful-stream-processing-in-apache-spark-streaming
- Getting Started with Apache Kafka: Ryan Plant
  - pluralsight.com/courses/apache-kafka-getting-started
- Real-Time Data Pipelines with Spark, Kafka, and Cassandra (on Docker): Nanda Vijaydev
  - bluedata.com/blog/2016/02/real-time-data-pipelines-spark-kafka-cassandra-on-docker/
- The world beyond batch: Streaming 101/102: Tyler Akidau
  - https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-101 / 102

Summary

- Competition
- Deeper Understanding
- Stateful Streams
- Maintenance