Introduction to Machine Learning on Apache Spark

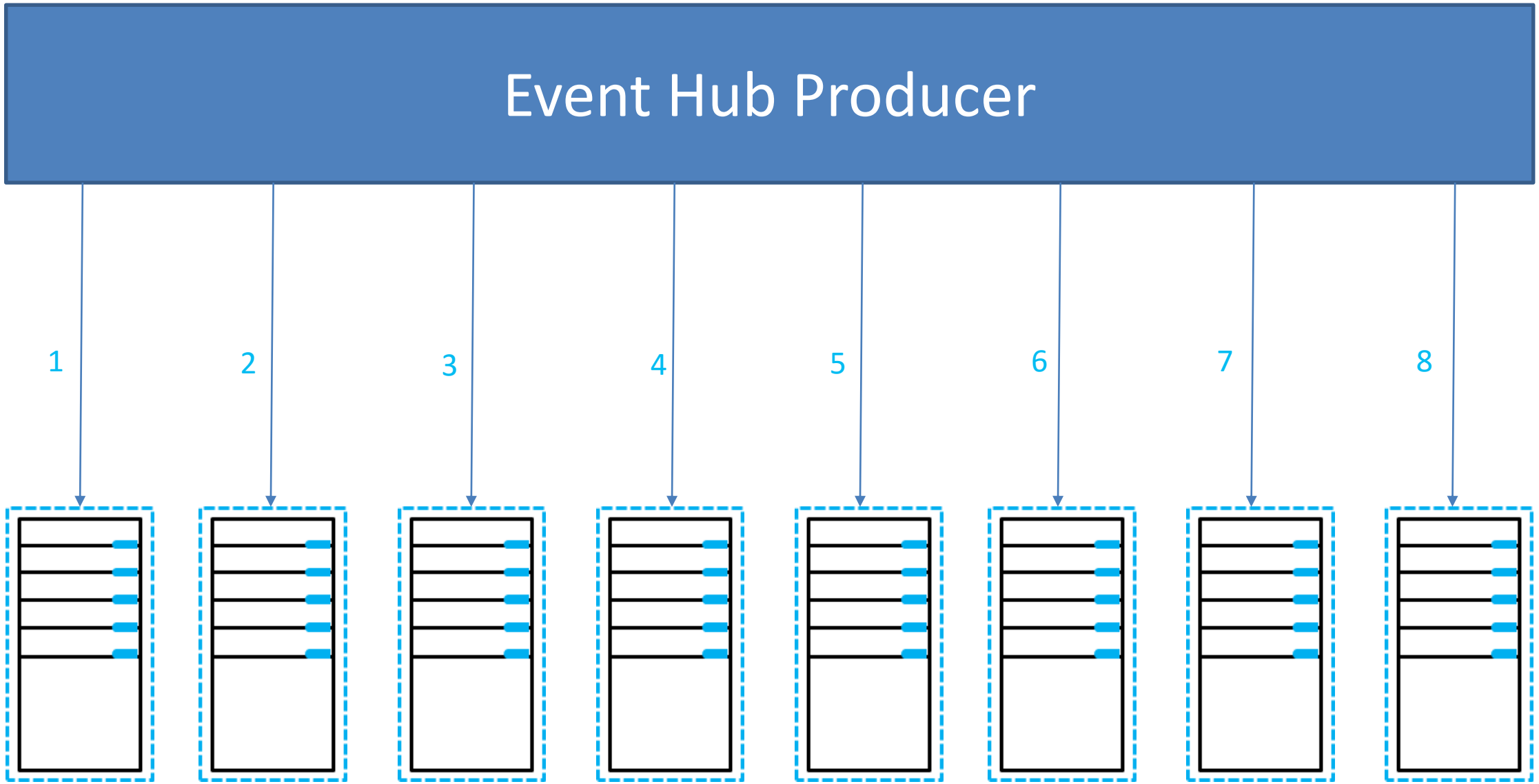
03 | Machine Learning with Streaming

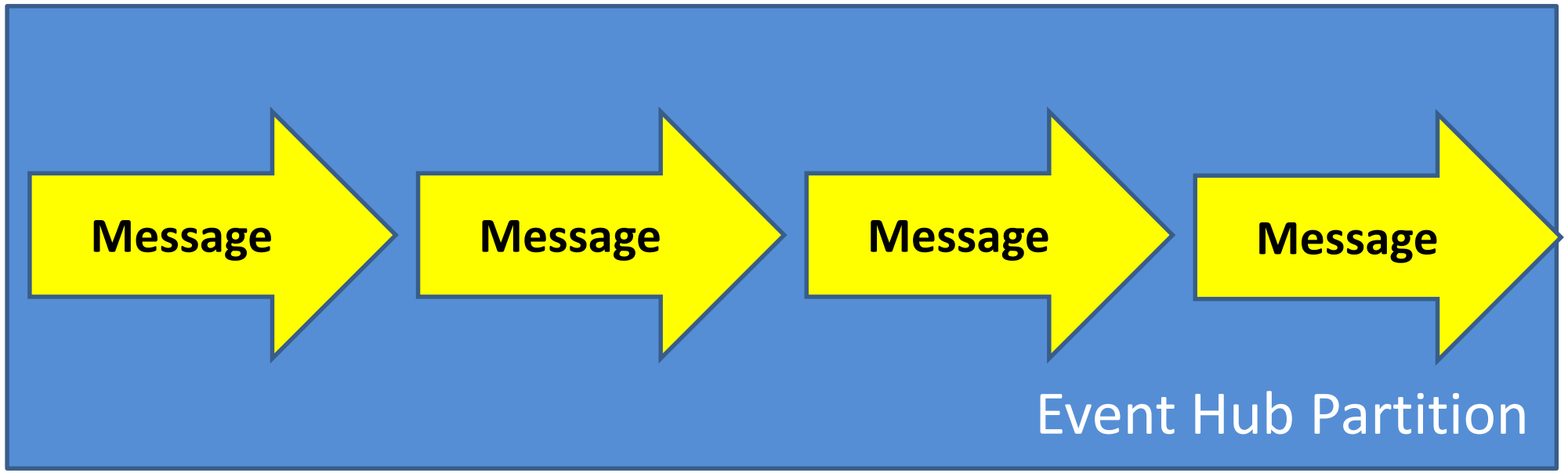
Richard Conway | Microsoft Azure MVP, Elastacloud

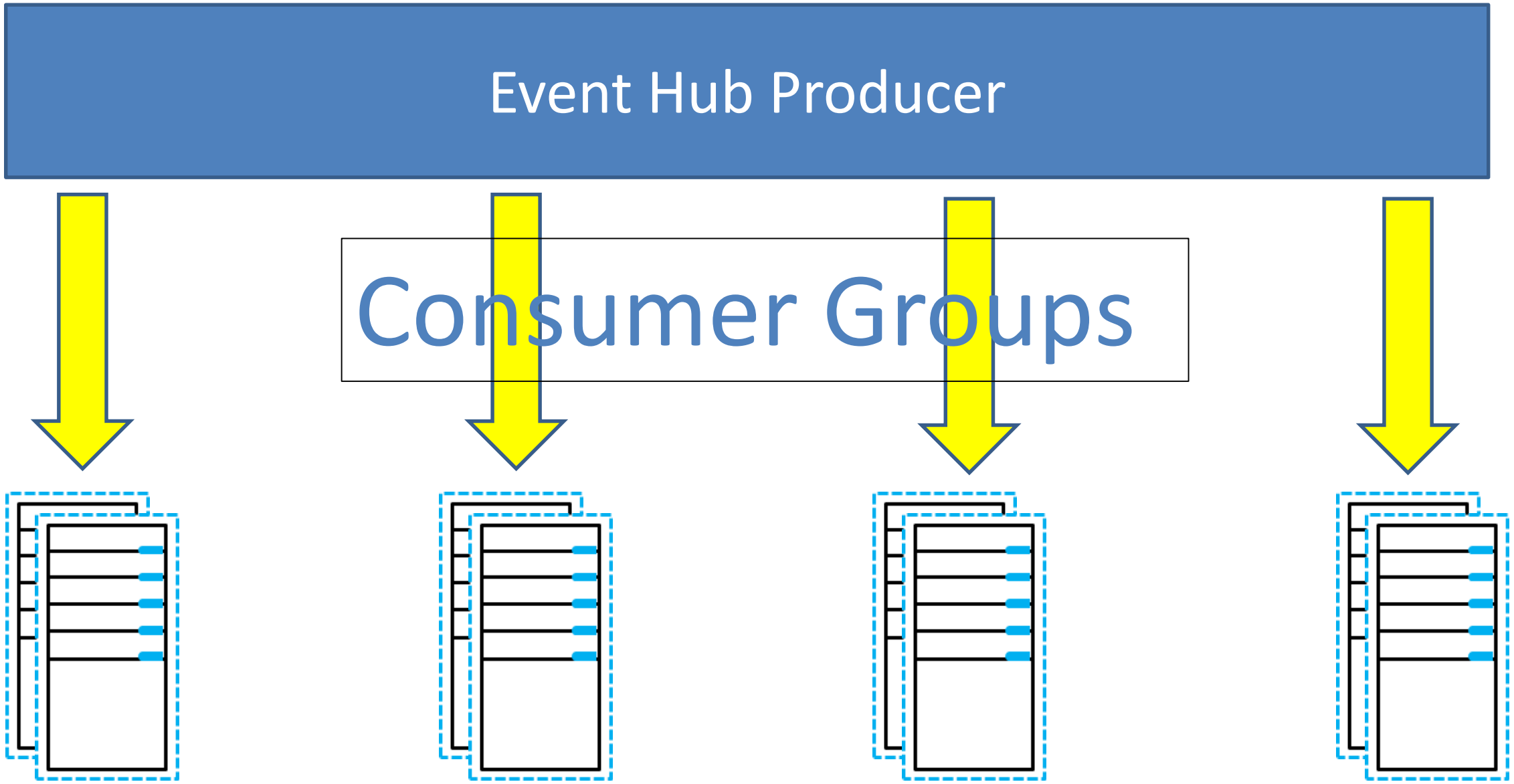
- Working with the Event Hub
- How does Spark Streaming work?
- How do I use DStreams and sliding windows?
- How do I use DataFrames to manipulate data?
- How do I use Machine Learning with a text input?
- How do you build a Machine Learning state machine?
- How do you Stream file updates to MLLib?
- How do I use time series analysis with Spark Streaming?

Working with the Event Hub

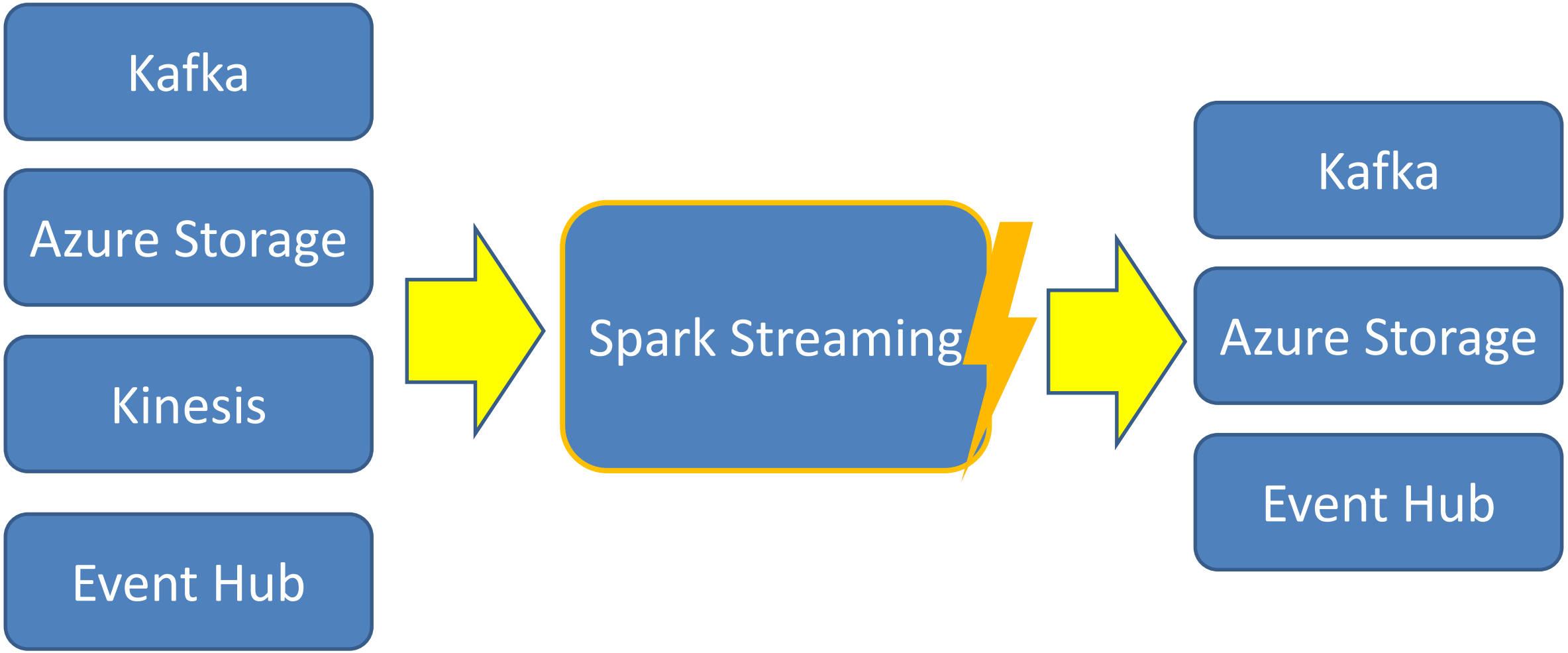
- Process millions of messages into Azure
- Time-based event buffering
- Perform real-time analytics in Azure using a variety of mechanisms
- Supports HTTP sending of messages
- Supports Advanced Message Queuing Protocol (AMQP)
- Scale clients with consumer groups
- Manage scale through partitions


Event Hub Producer
1
2
3
4
5
6
7
8

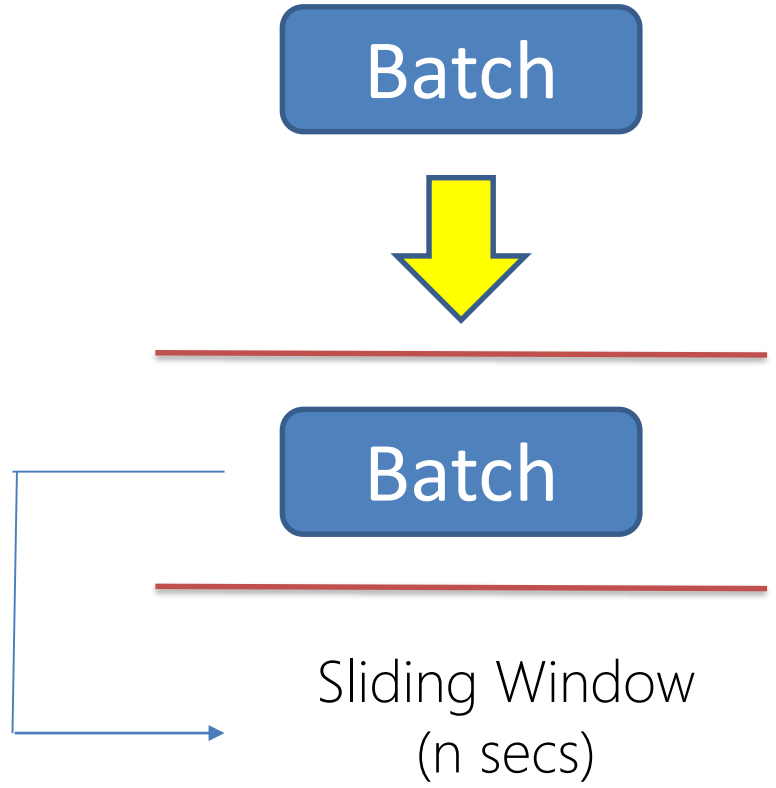

Message
Message
Message
Message
Event Hub Partition


Event Hub Producer
Consumer Groups

How does Spark Streaming work?

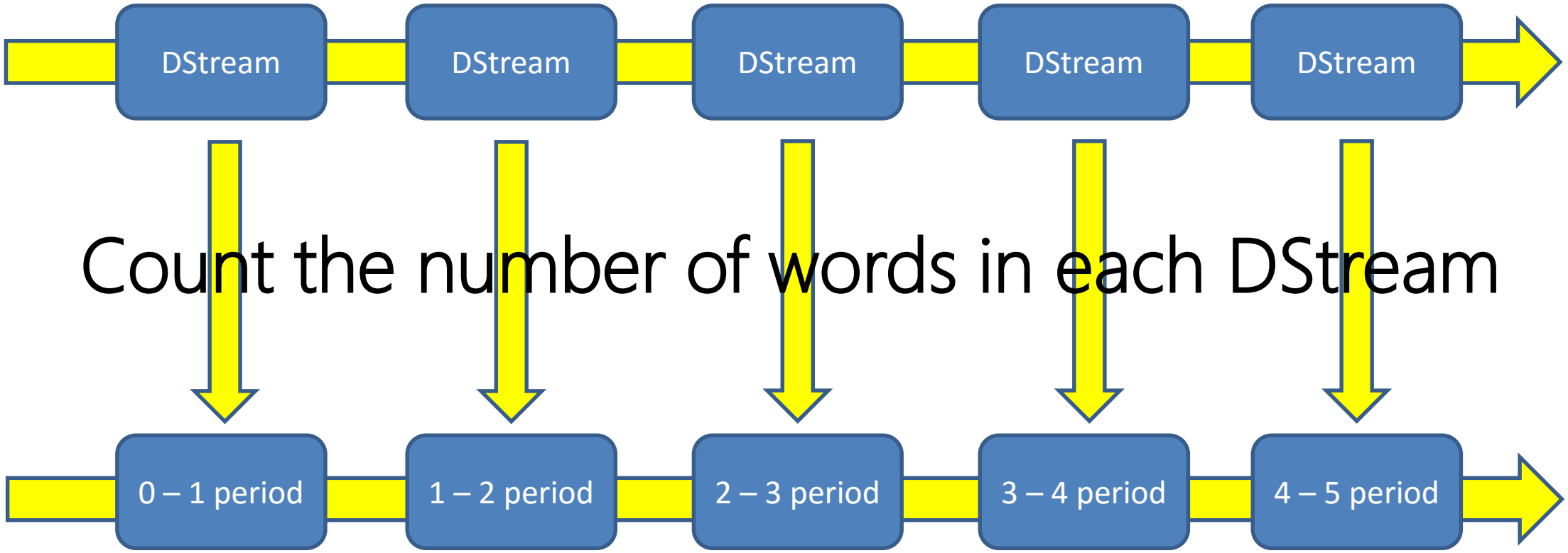

Kafka
Azure Storage
Kinesis
Event Hub
Spark Streaming
Kafka
Azure Storage
Event Hub

- Spark Streaming processes data in micro-batches
- The “streams” can be received across many nodes
- You enable a “Receiver” on each data node to receive messages
- Messages received in sliding windows

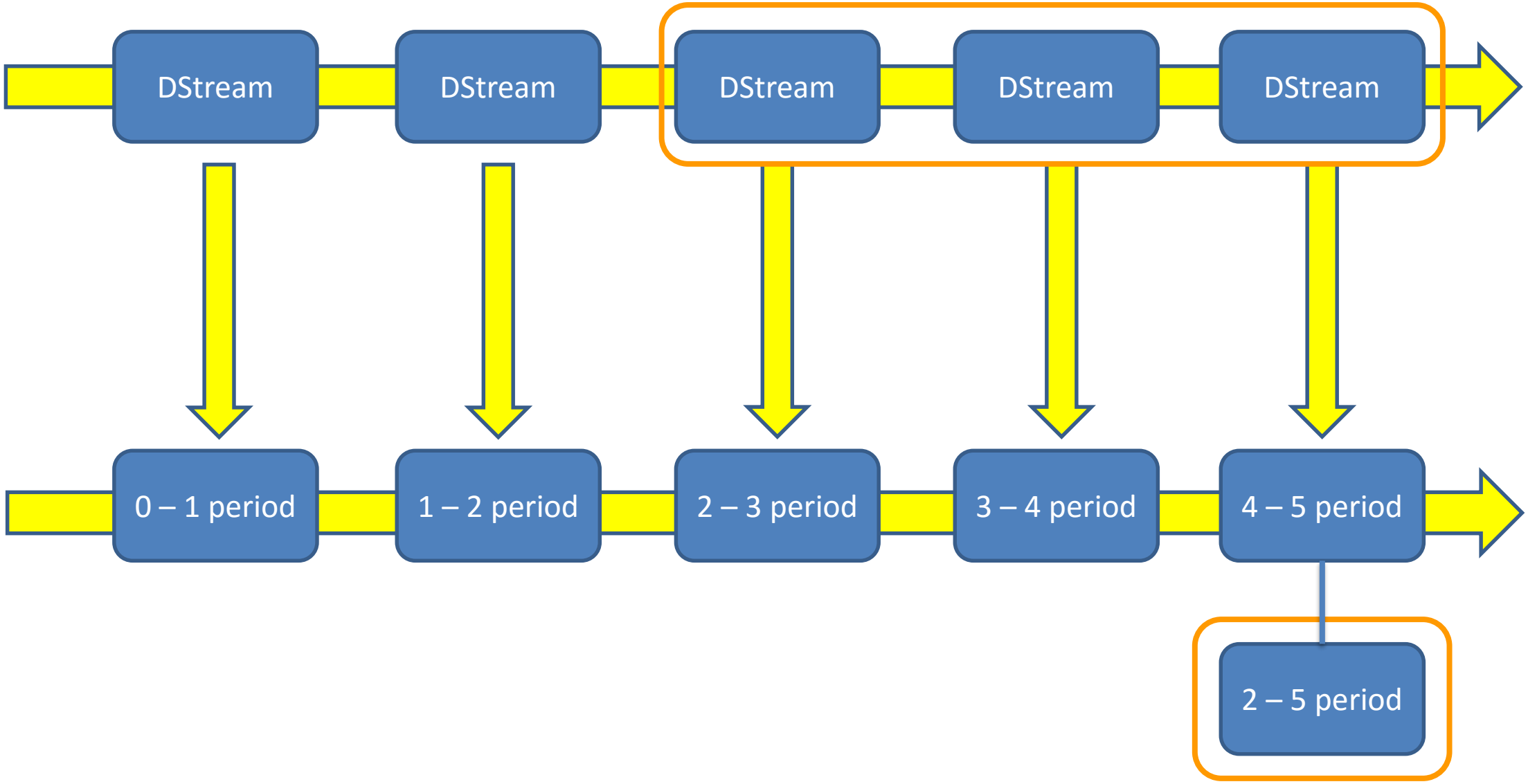


```
val sparkConfiguration = new SparkConf().setAppName("edX Courses are the best!")
val sparkContext = new SparkContext(sparkConfiguration)

val streamingContext = new StreamingContext(sparkContext, Seconds(5))
streamingContext.checkpoint("/usr/checkpoint"))

val eventHubsStream = EventHubsUtils.createUnionStream(streamingContext,
eventHubsParameters)

val eventHubsWindowedStream = eventHubsStream.window(Seconds(10))

val batchEventCount = eventHubsWindowedStream.count()
batchEventCount.print()

streamingContext.start()
streamingContext.awaitTermination()
```

How do I use DStreams and sliding windows?

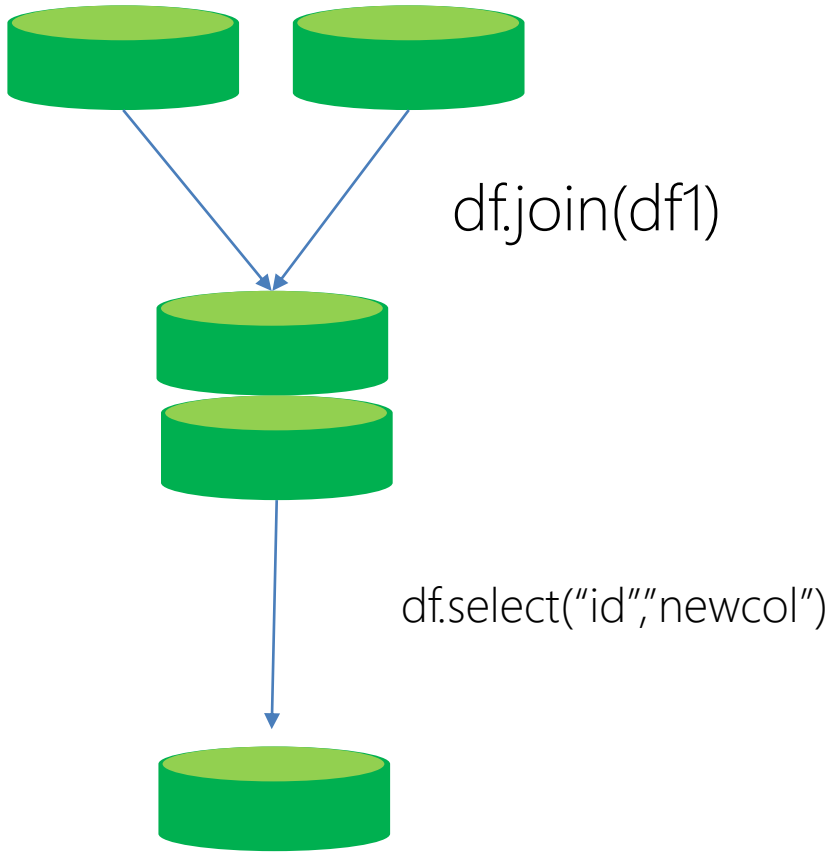


Count the number of words in each DStream


DStream
DStream
DStream
DStream
DStream
0 – 1 period
1 – 2 period
2 – 3 period
3 – 4 period
4 – 5 period
2 – 5 period

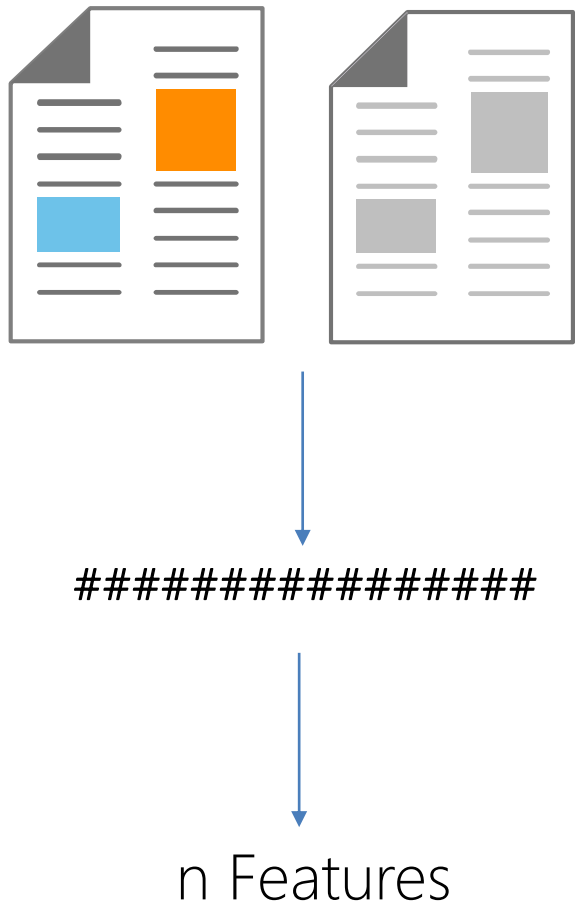
How do I use DataFrames to manipulate data?

- Perform joins between DataFrames
- Enrich DataFrames with columns from other DFs
- Execute UDFs on DataFrames
- Build in continuous aggregation using streaming
- All kinds of aggregations possible
- Persistence and caching



How do I use Machine Learning with a text input?

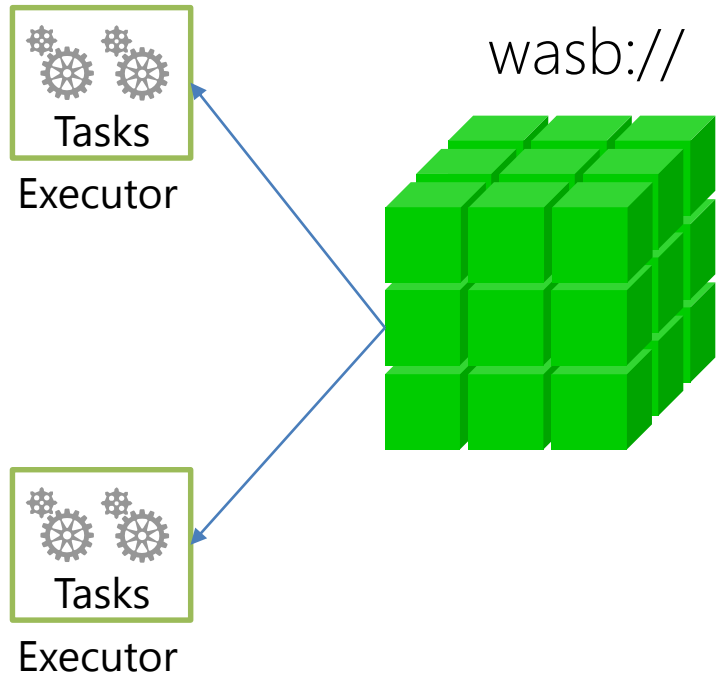
- Uses a document corpus as input
- Support TF-hashing
- Supports IDF as well
- Separated so can be used together or apart
- Supported independently or through pipelines
- Support word2vec



```
import org.apache.spark.mllib.feature.HashingTF
import org.apache.spark.mllib.feature.IDF

val documents = sc.textFile("/tweets").map(_.split(" ").toSeq

val hashingTF = new HashingTF()
val tf = hashingTF.transform(documents)
val idf = new IDF().fit(tf)
val tfidf = idf.transform(tf)
```

How do you stream file updates to MLLib?

- Direct streaming integration from MLLib
- Streams all new files from HDFS (WASB) or local file system
- No streaming (Receiver) code needed
- Supports Linear Regression and KMeans



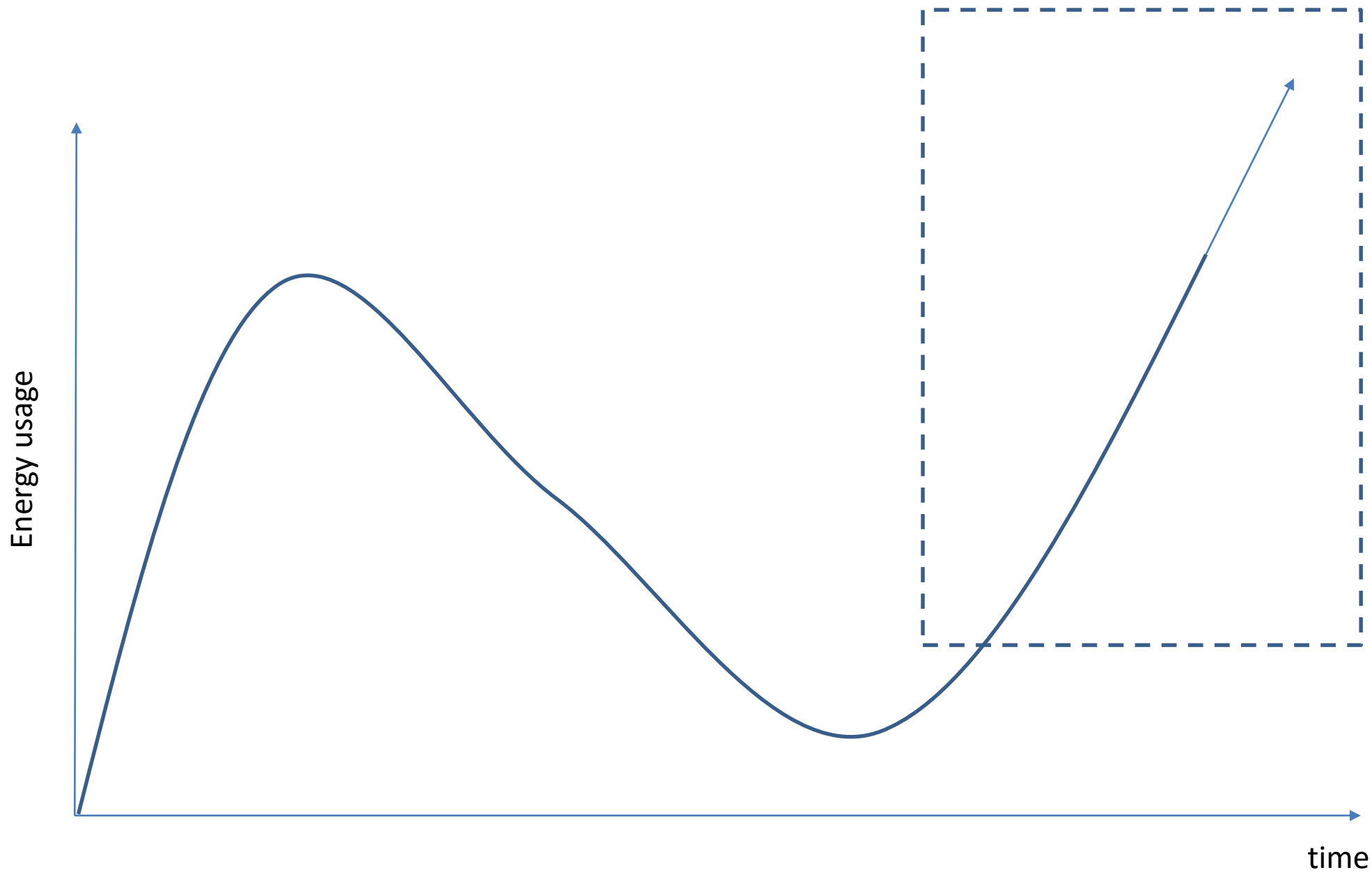
```
val train = ssc.textFileStream("wasb:///iris-train").map(Vectors.parse)
val test = ssc.textFileStream("wasb:///iris-test").map(LabeledPoint.parse)

val model = new StreamingKMeans()
  .setK(2)
  .setDecayFactor(1.0)
  .setRandomCenters(3, 0.0)

model.trainOn(train)
model.predictOnValues(train.map(lp => (lp.label, lp.features))).print()

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

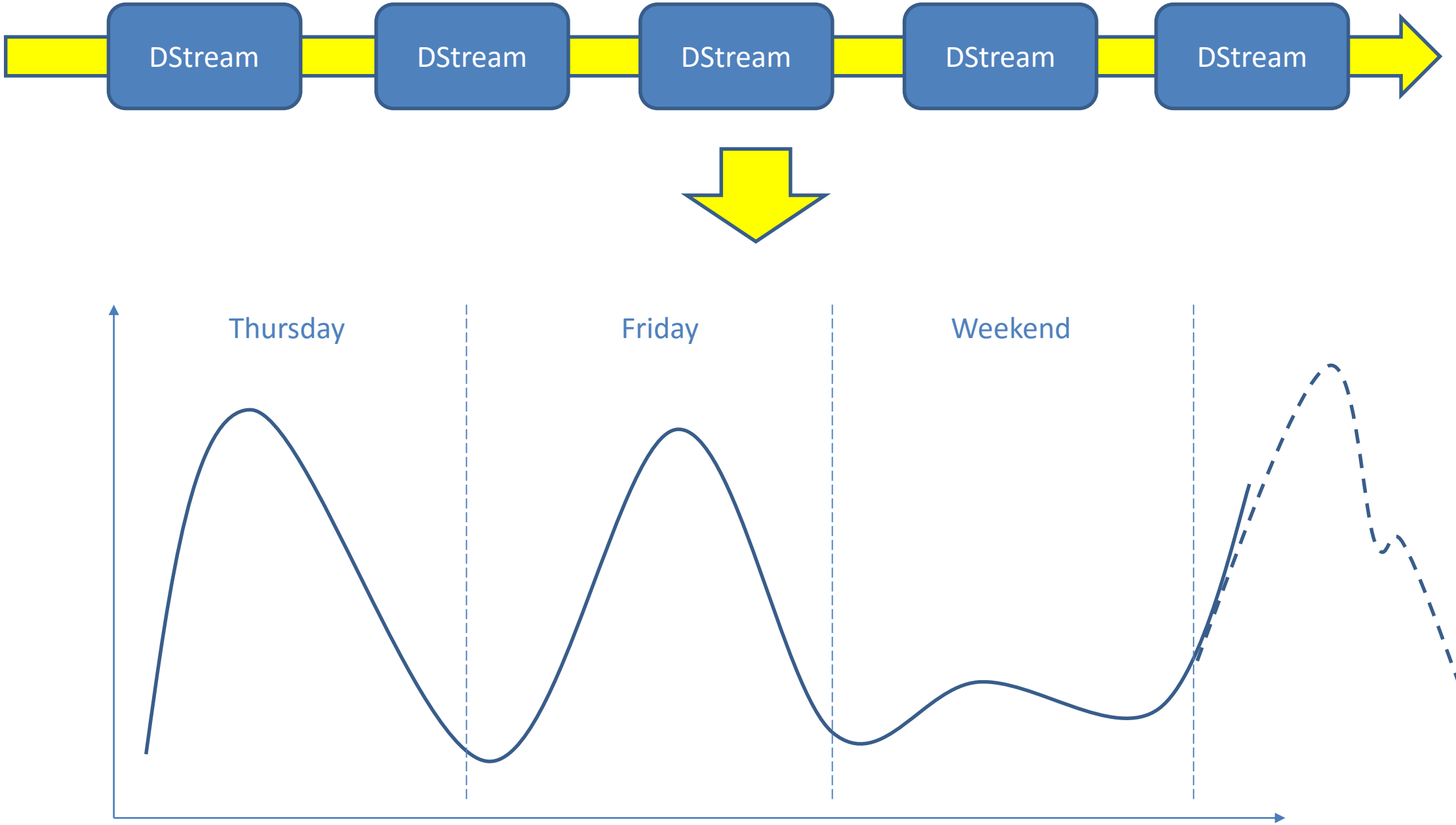
DEMO

Streaming Machine Learning with Wasb

How do I use time series analysis with Spark Streaming?

- Time series analysis is a very useful statistical technique
- It looks at continuous ranges of time
- It removes the “periodicity” of the series
- It allows for a calculation of a “Moving Average” to highlight changes
- With Time Series you can check serial correlation
- You can make predictions on the future of the series by “featurizing” related details

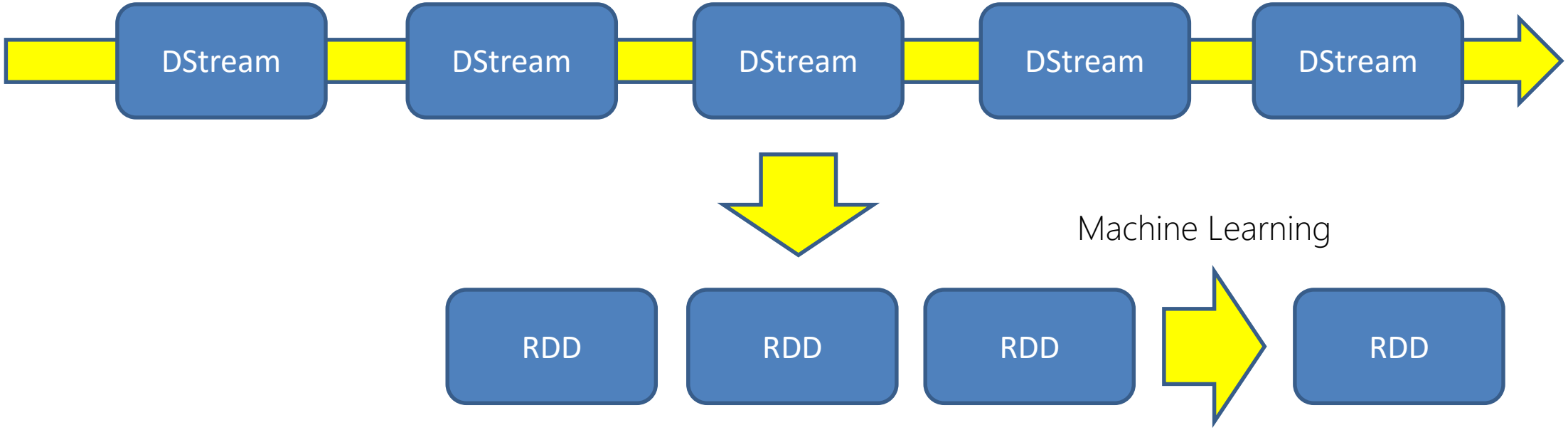

Energy usage
time


DStream
DStream
DStream
DStream
DStream
Thursday
Friday
Weekend

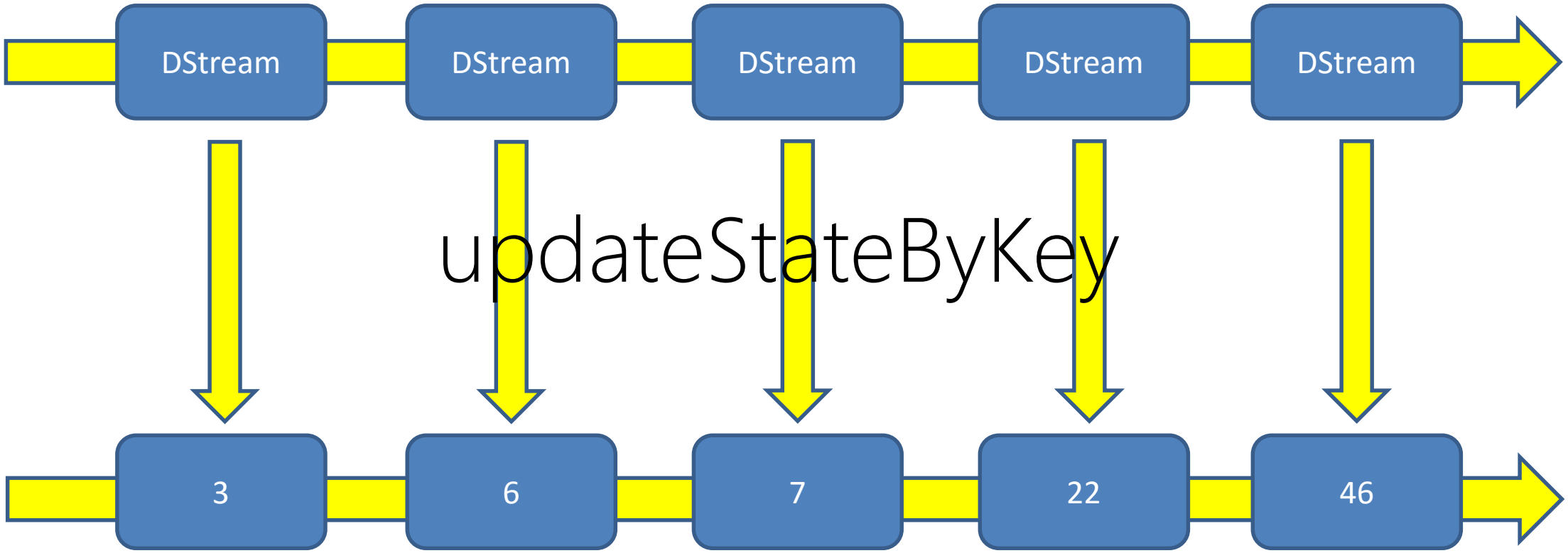
```
val eventHubsWindowedStream =
eventHubsStream.window(Minutes(10))
val deviceObs= loadDeviceData(eventHubsWindowedStream)

dtIndex = DateTimeIndex.uniform(
DateTime("2016-02-03"), new DateTime("2016-03-03"), new
BusinessDayFrequency(1))

val drdd = TimeSeriesRDD.timeSeriesRDDFromObservations(dtIndex,
deviceObs, "timestamp", "devicetype", "voltage")

drdd.cache()
val filled = drdd.fill("linear")
```

- There are several usage patterns with Spark Streaming and machine learning
  - Time series analysis
  - DStream to RDD message inspection
  - UpdateStateByKey functions
  - DStream transforms and predictive enrichment

How do you build a Machine Learning state machine?




```
dstream.foreachRDD { rdd =>
  rdd.foreach { record =>
    val anom = checkForAnomaliesViaKMeans(record)
    sendEventHubMessage(anom)
  }
}
```


DStream
DStream
DStream
DStream
DStream
updateStateByKey
3
6
7
22
46

```
def updateProbability(key: Int, ticksPlusValues: (long, Seq[Int]), probability:
Option[Int]): Option[Int] = {
    val newCount = getNewProbability(ticksPlusValues._1,
runningCount.getOrElse(0) + ticksPlusValues._2.sum)
    Some(newCount)
}

val runningProbability = pairs.updateStateByKey[Int](updateProbability _)
```

- Working with the Event Hub
- How does Spark Streaming work?
- How do I use DStreams and sliding windows?
- How do I use DataFrames to manipulate data?
- How do I use Machine Learning with a text input?
- How do you build a Machine Learning state machine?
- How do you Stream file updates to MLLib?
- How do you execute code remotely in Apache Spark?
- How do I use time series analysis with Spark Streaming?

Microsoft