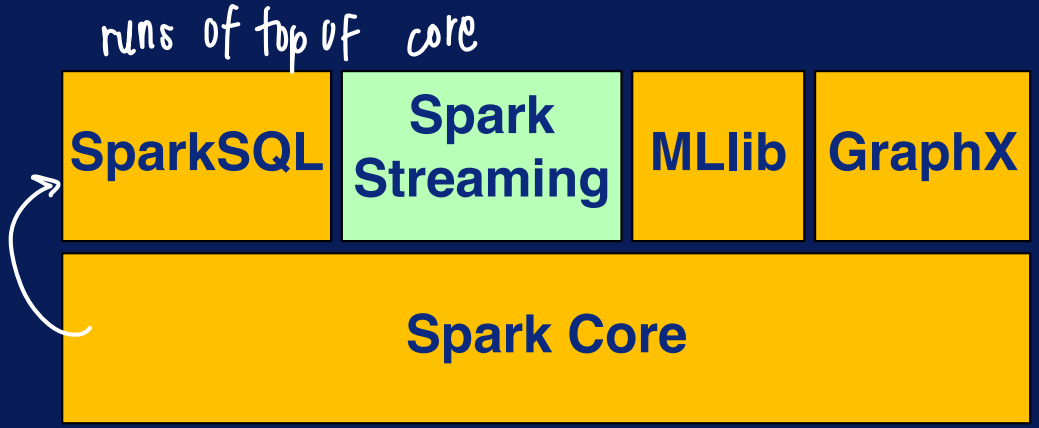
Spark Streaming

After this video you will be able to..

- Summarize how Spark reads streaming data
- List several sources of streaming data supported by Spark
- Describe Spark's sliding windows



Spark Streaming

- Scalable processing for real-time analytics (data)
- Data streams converted to discrete RDDs (or grouped) ⇝ which can then be processed in parallel
- Has APIs for Scala, Java, and Python

Spark Streaming Sources

can read data from many different types of resources

- Kafka : high throughput published subscribed messaging system
- Flume : collects and aggregates log data
- HDFS
- S3

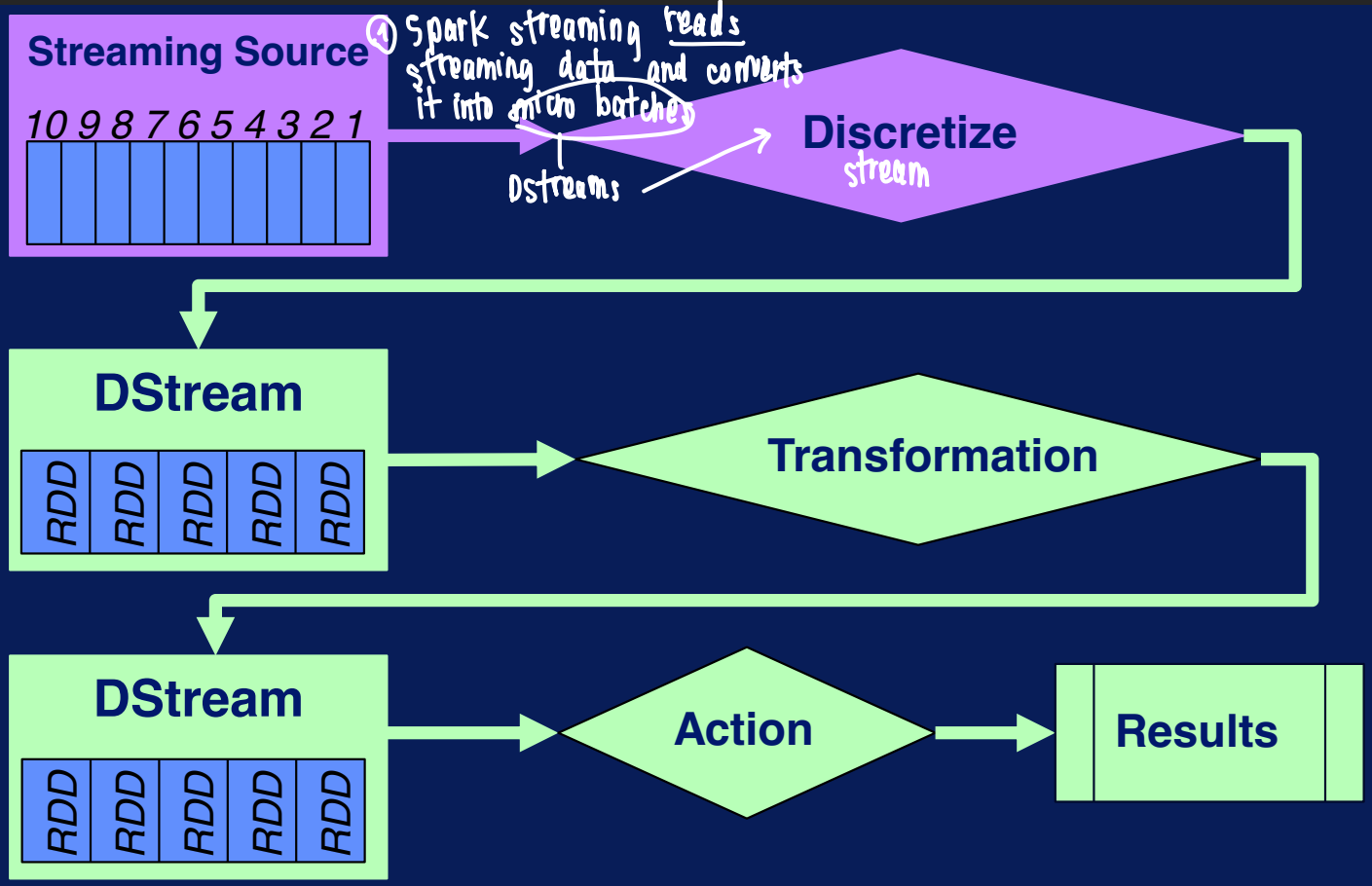
can read from batch spark streaming and other no sql databases

- Twitter
- Sockets (raw TCP)

can read from

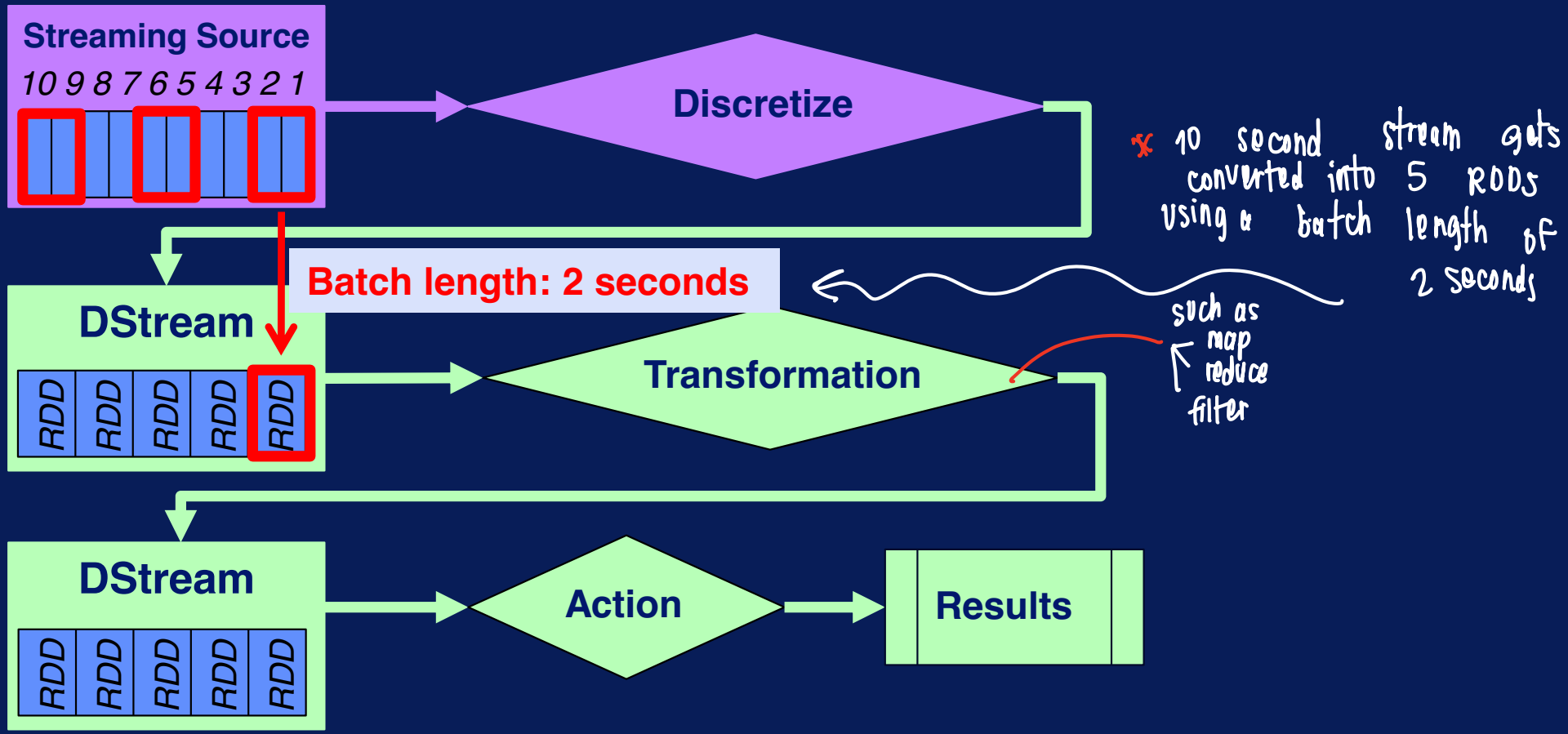
- …etc. ~ other data sources that are real-time data providers

Creating and Processing DStreams



flow of actions and transformations

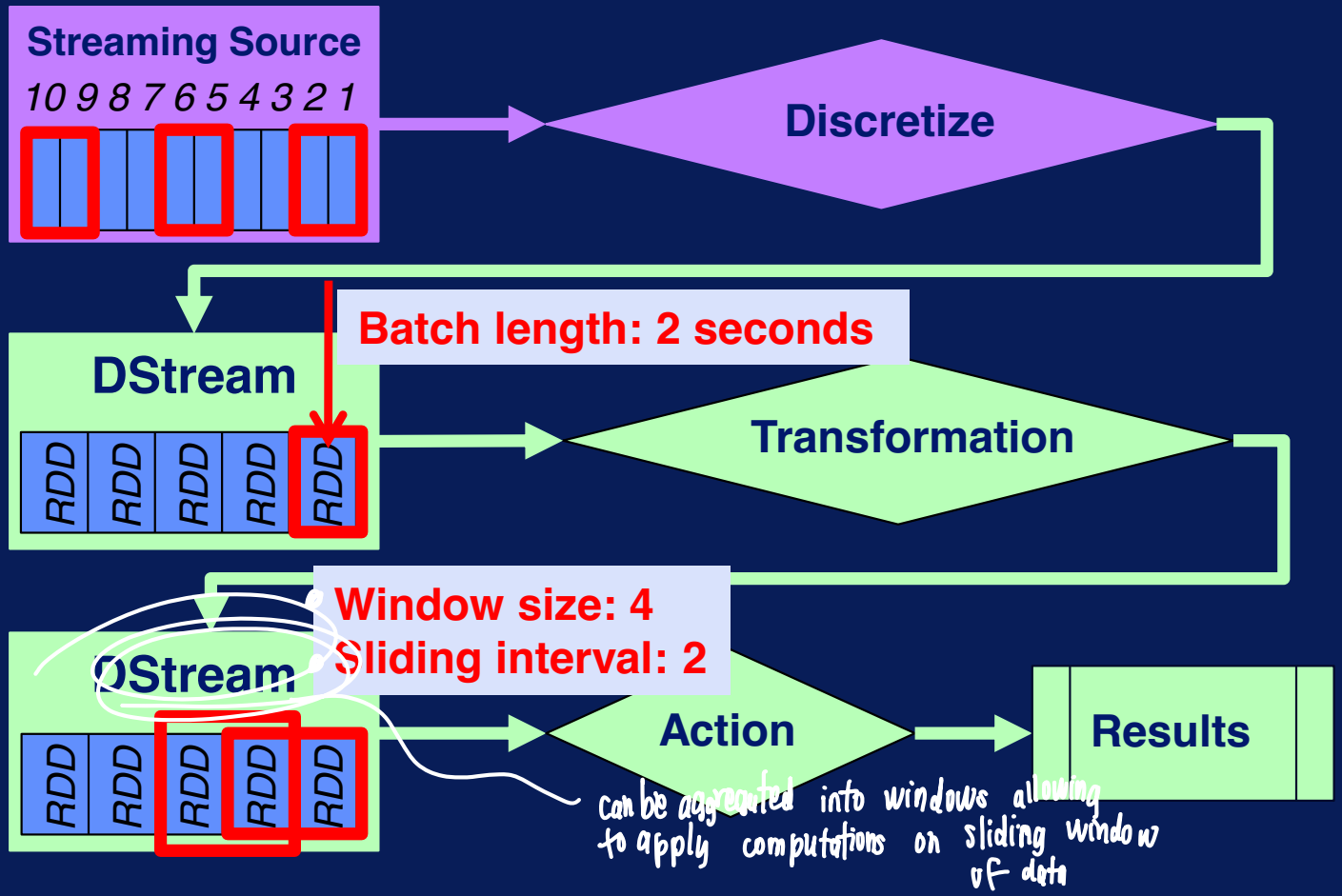
Creating and Processing DStreams



10 second stream gets converted into 5 RDDs using a batch length of 2 seconds

such as map reduce filter

Creating and Processing DStreams



DStream (Discretized stream): discrete packets of RDD generated from stream source

Main Take-Aways

Spark streaming: Spark's library to work with streaming data in near real time

- Spark uses DStreams to make discrete RDDs from streaming data.
  - Same transformations and calculations applied to batch RDDs can be applied
- DStreams can create a sliding window to perform calculations on a window of time.

can be used like any other RDD