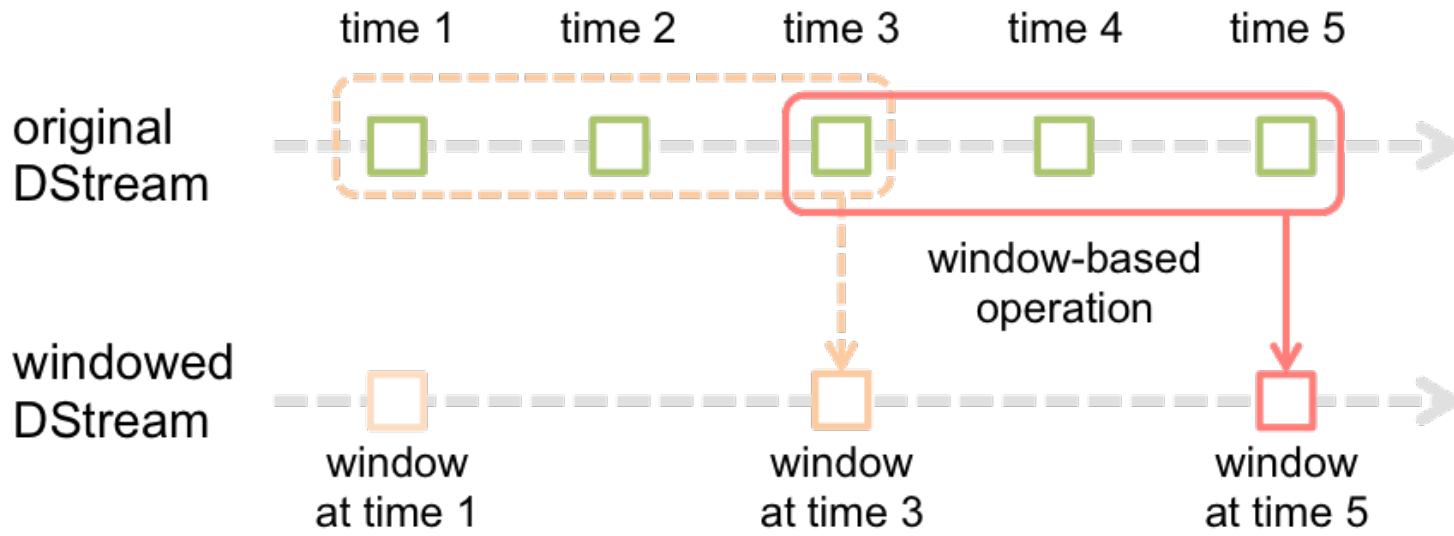
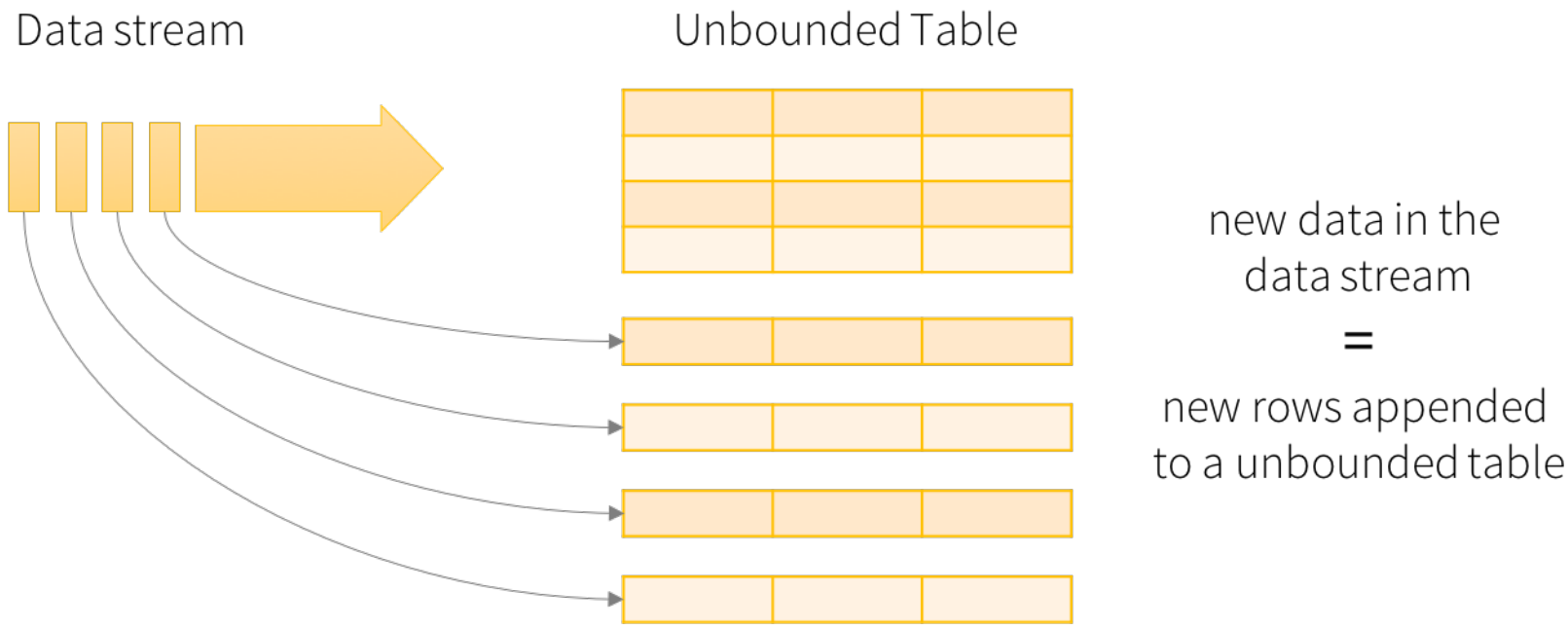

input data stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data


time 1
time 2
time 3
time 4
time 5
original DStream
window-based operation
windowed DStream
window at time 1
window at time 3
window at time 5

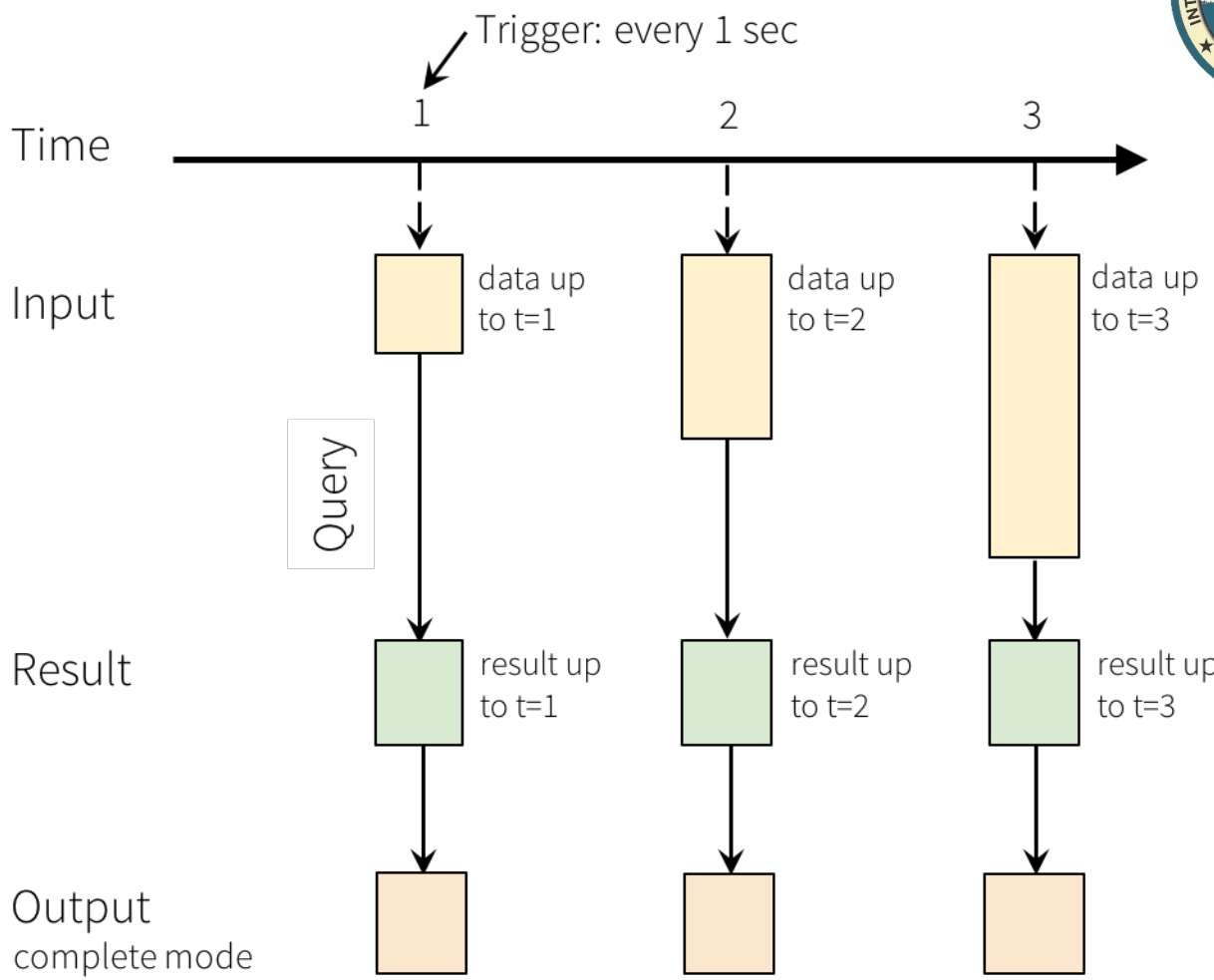
Structured Streaming

- It is built on the Spark SQL engine without the user having to reason about streaming.
  - You can express your streaming computation the same way you would express a batch computation on static data.
  - The Spark SQL engine will take care of running it incrementally and continuously and updating the final result as streaming data continues to arrive.
- You can use the Dataset/DataFrame API in Scala, Java, Python or R

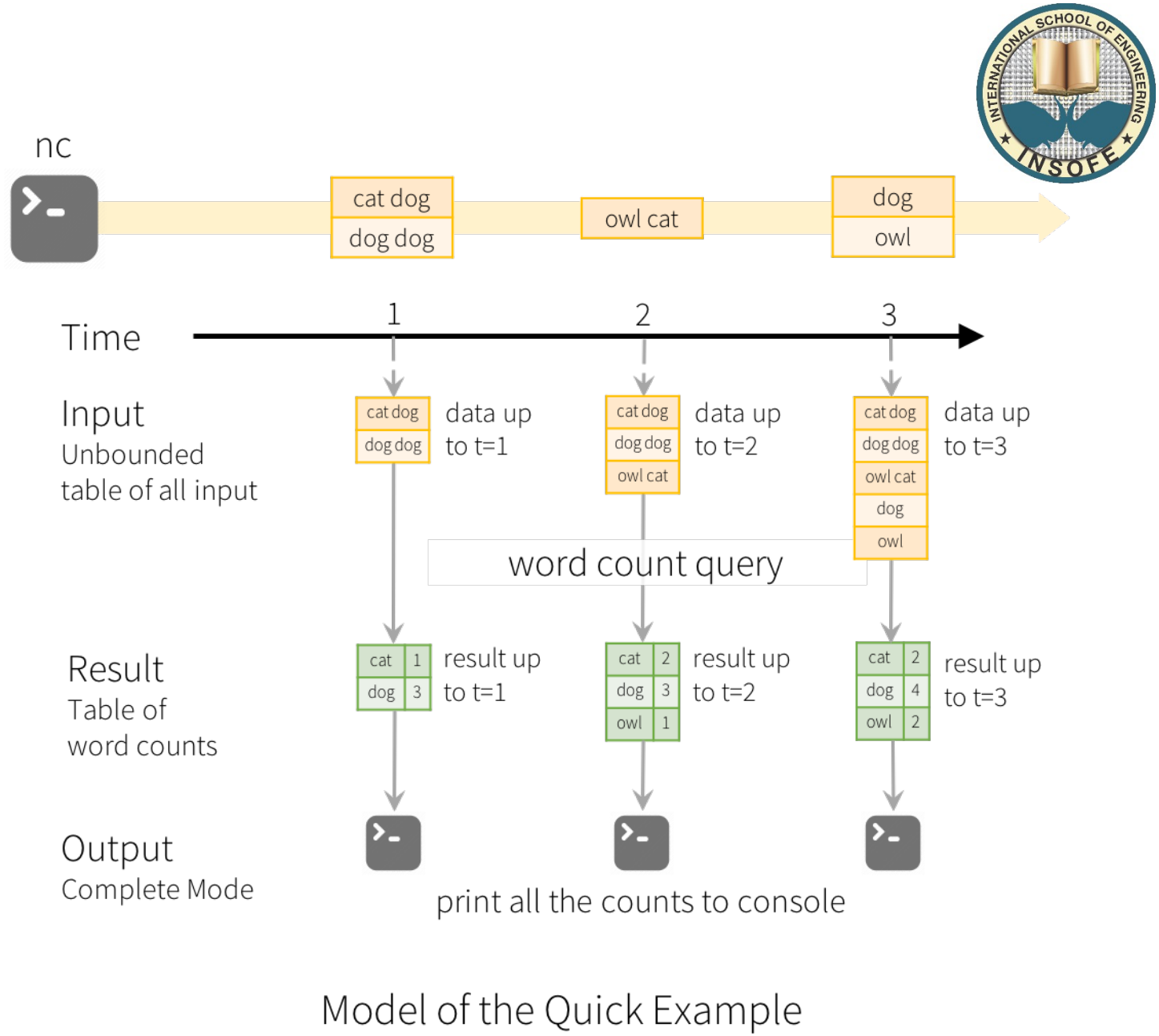
Quick Example



Data stream as an unbounded table

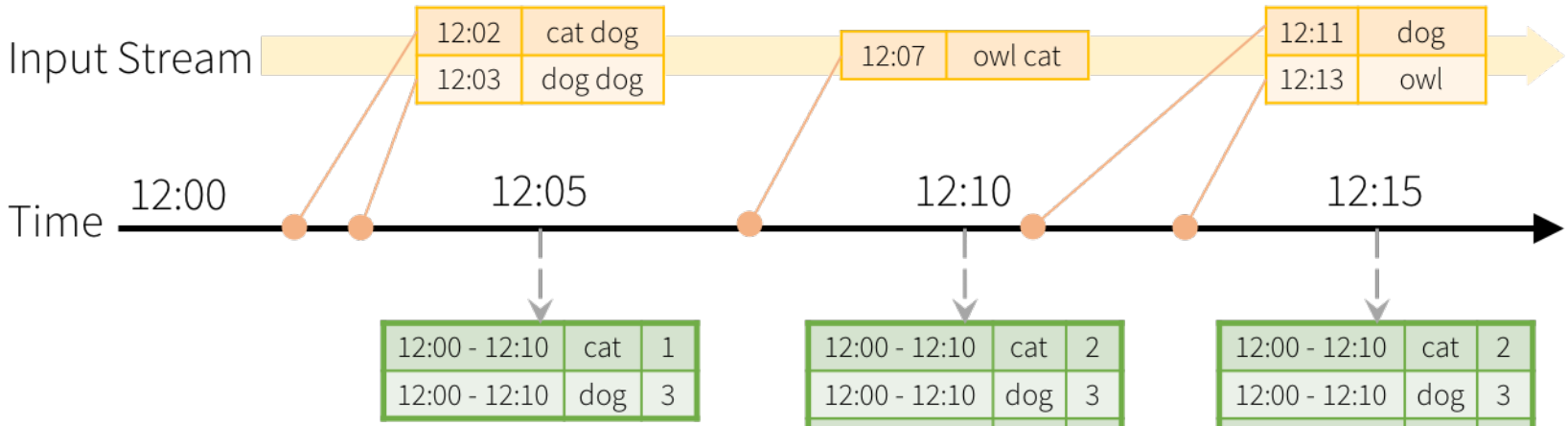


Programming Model for Structured Streaming



Model of the Quick Example

Window Operations



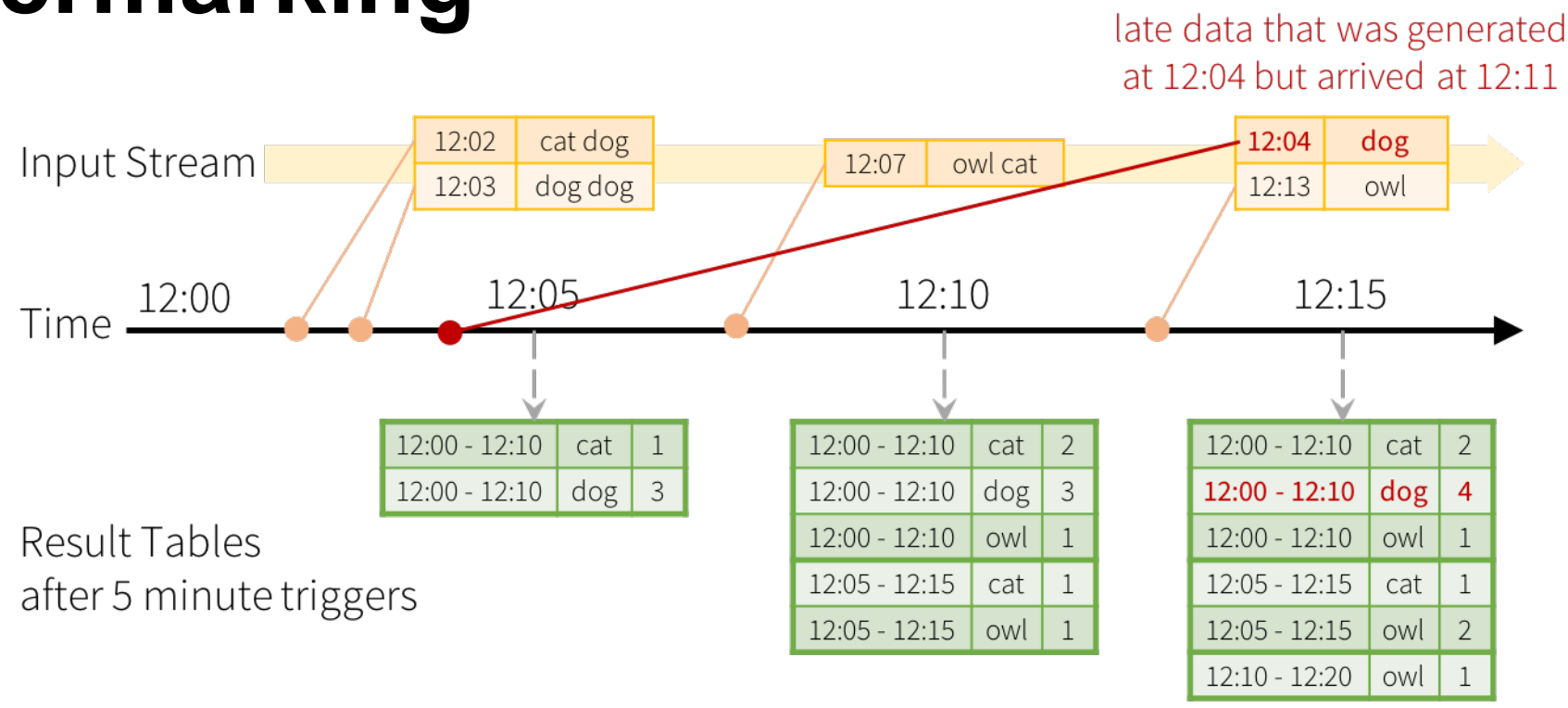
Result Tables
after 5 minute triggers

counts incremented for windows
12:00 - 12:10 and 12:05 - 12:15

counts incremented for windows
12:05 - 12:15 and 12:10 - 12:20

Windowed Grouped Aggregation
with 10 min windows, sliding every 5 mins

Handling Late Data and Watermarking



Late data handling in Windowed Grouped Aggregation