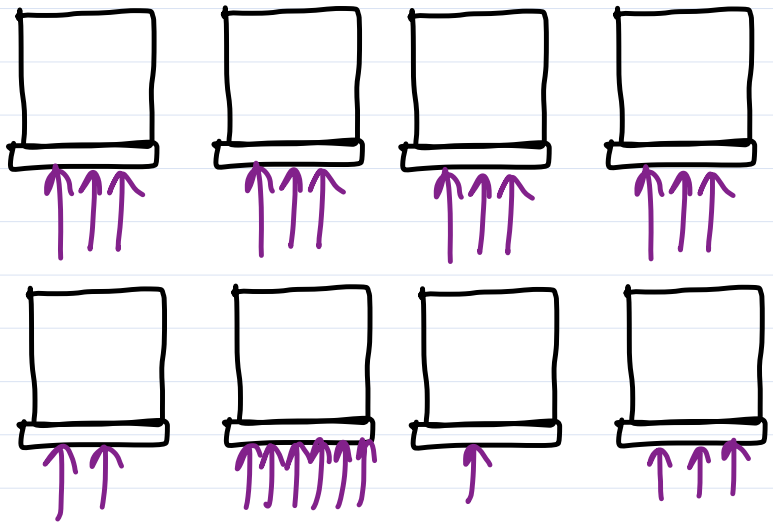
Big Data Technologies

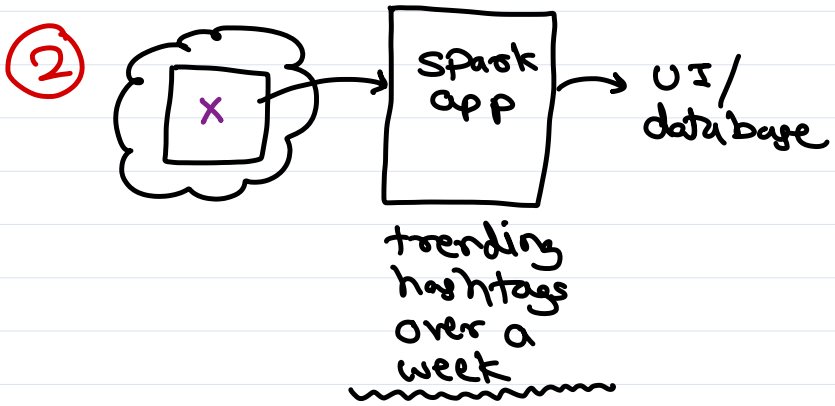
Trainer: Mr. Nilesh Ghule.

Stream Processing

① Load Balancing



②



huge state: hashtags of all tweets in a week.

③ exactly once processing

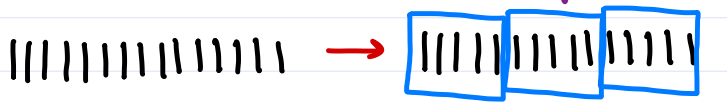
ⓐ at least once (may repeat)

ⓑ at most once (may skip)

ⓒ exactly once

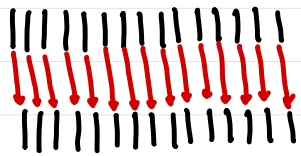
④ Low latency processing

ⓐ micro-batch approach



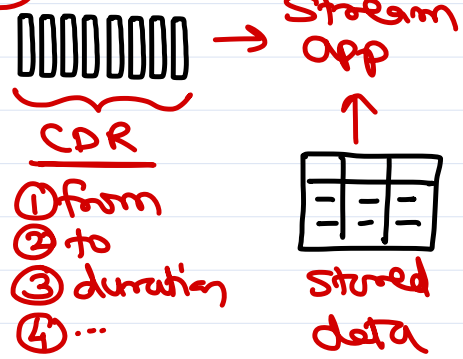
(secondy/minutes)

ⓑ Continuous processing

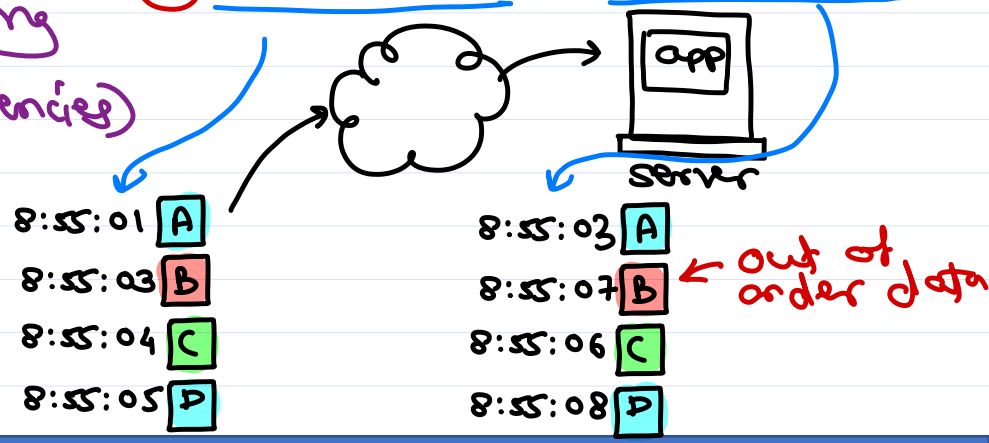


(lower latencies)

⑤ Join with external data

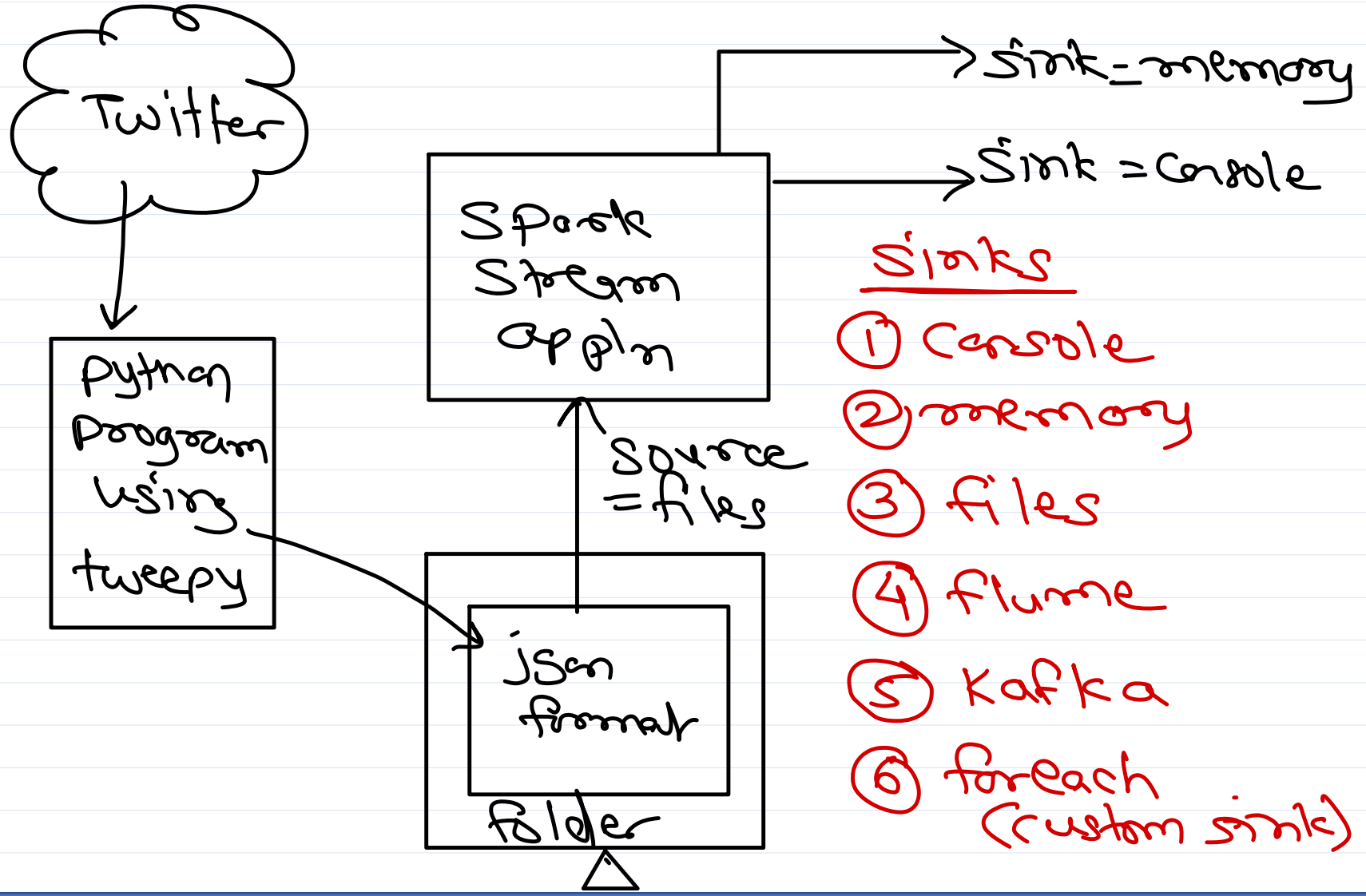


⑥ event time vs processing time



Sources

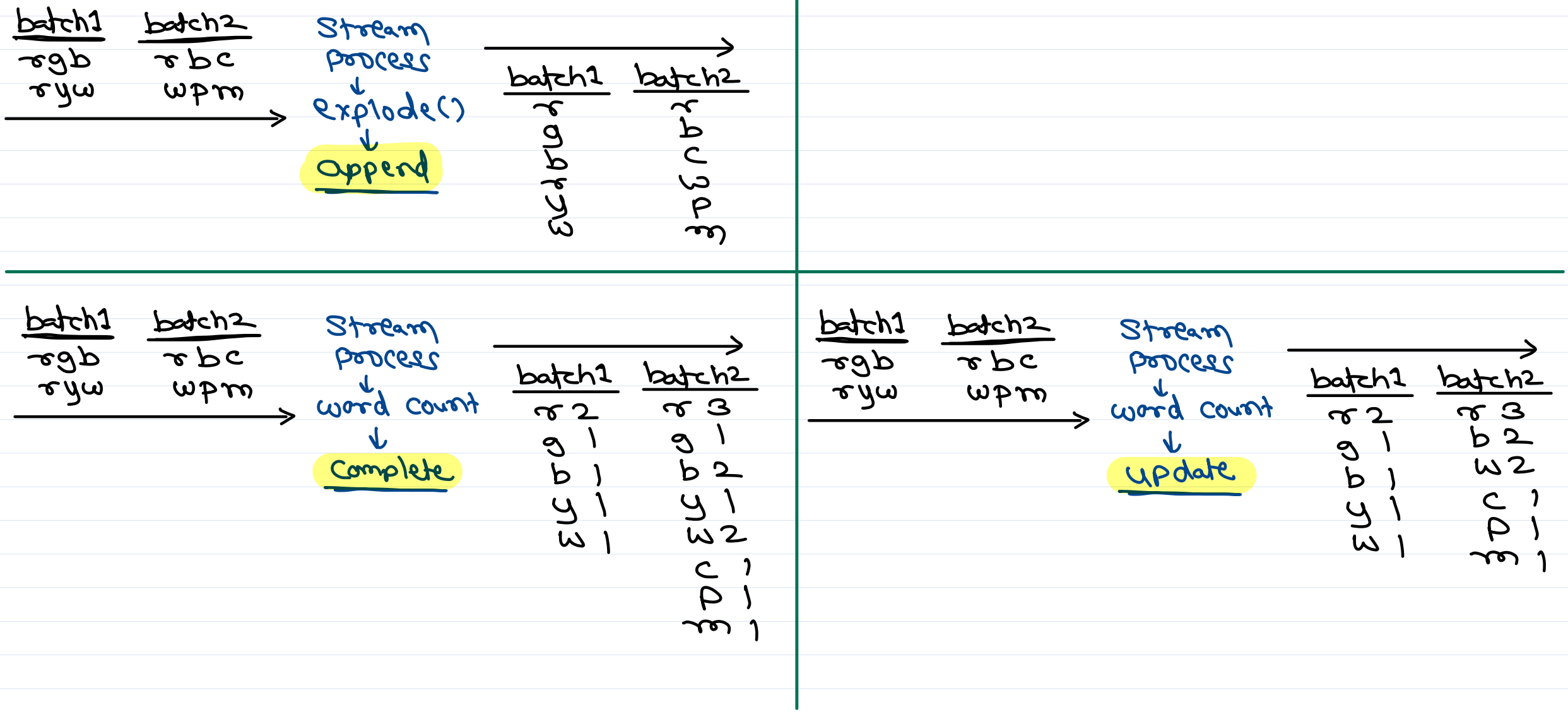
1. socket
2. files (csv, json, ...)
3. flume
4. kinesis
5. rate (testing)
6. kafka



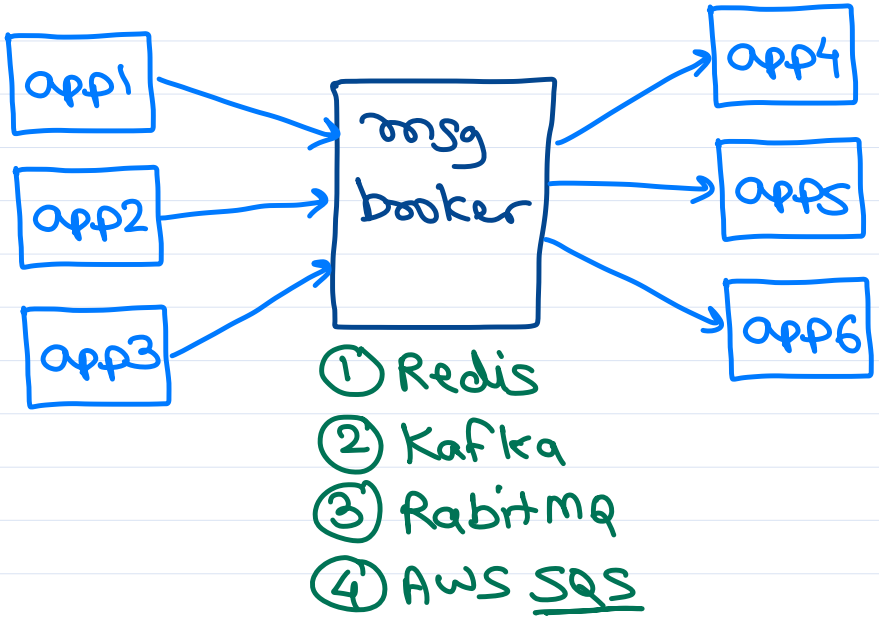
Sinks

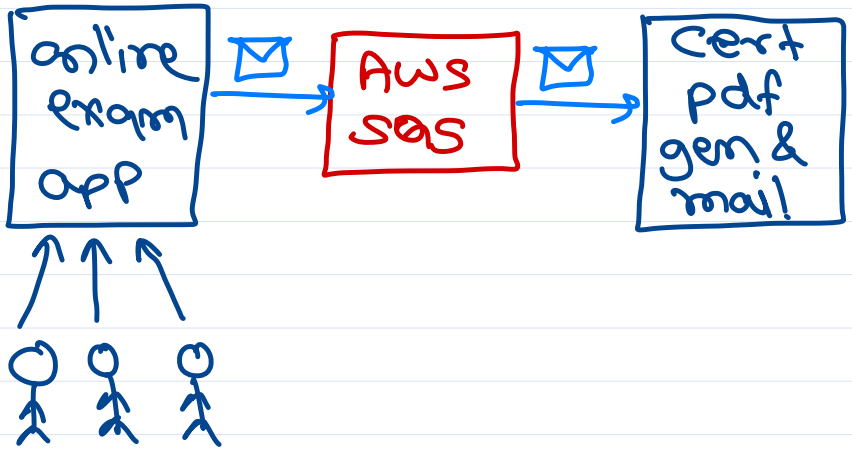
1. console
2. memory
3. files
4. flume
5. kafka
6. foreach (custom sink)

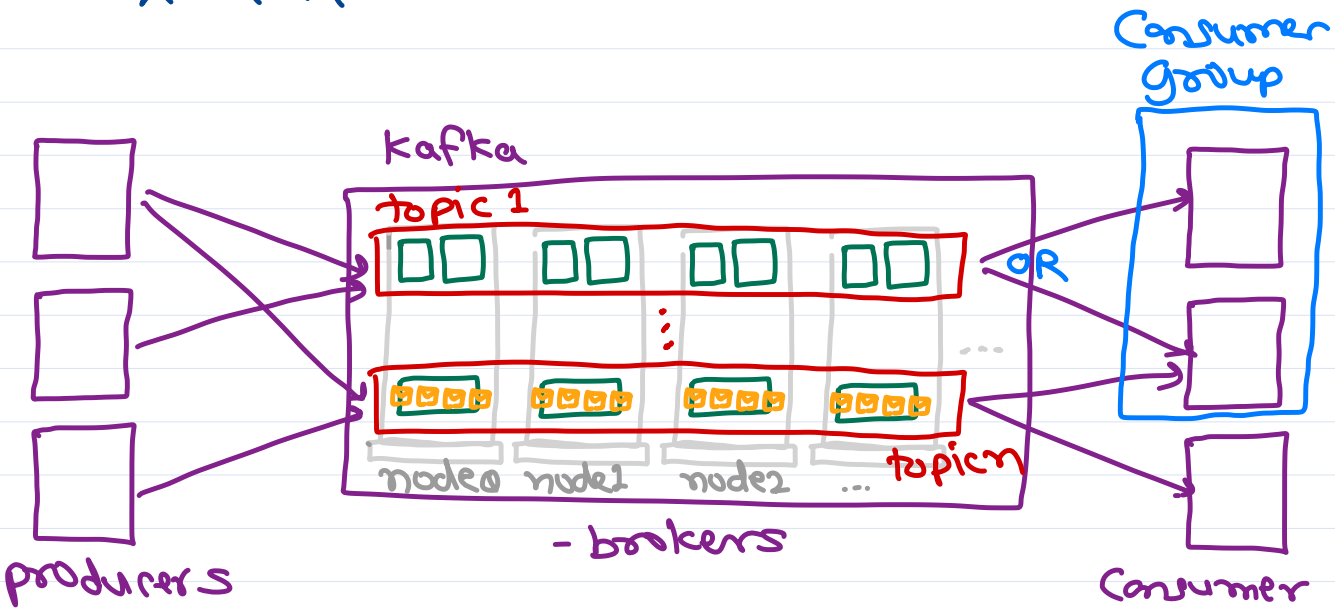
Stream Output Mode



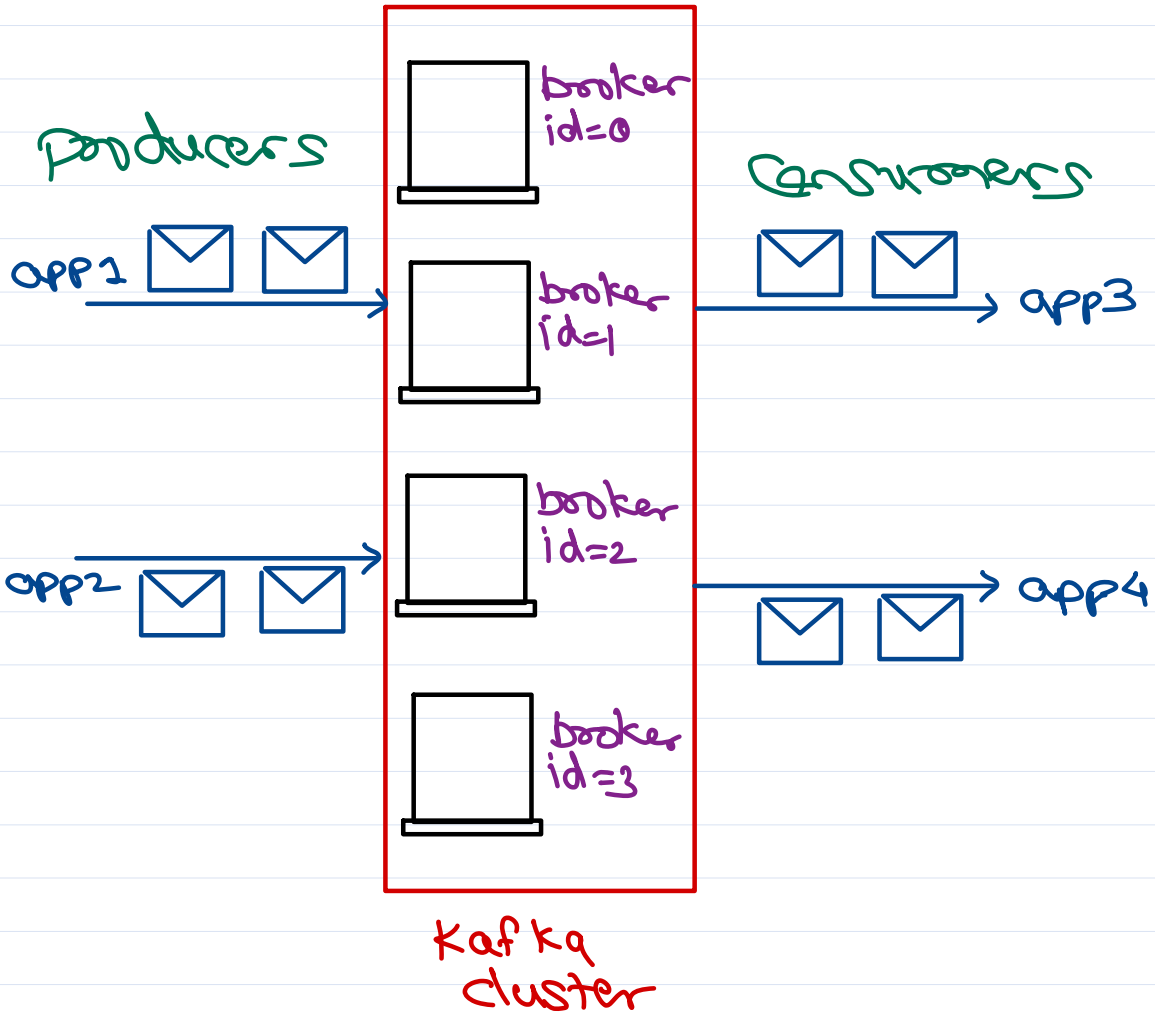
* Distributed message broker

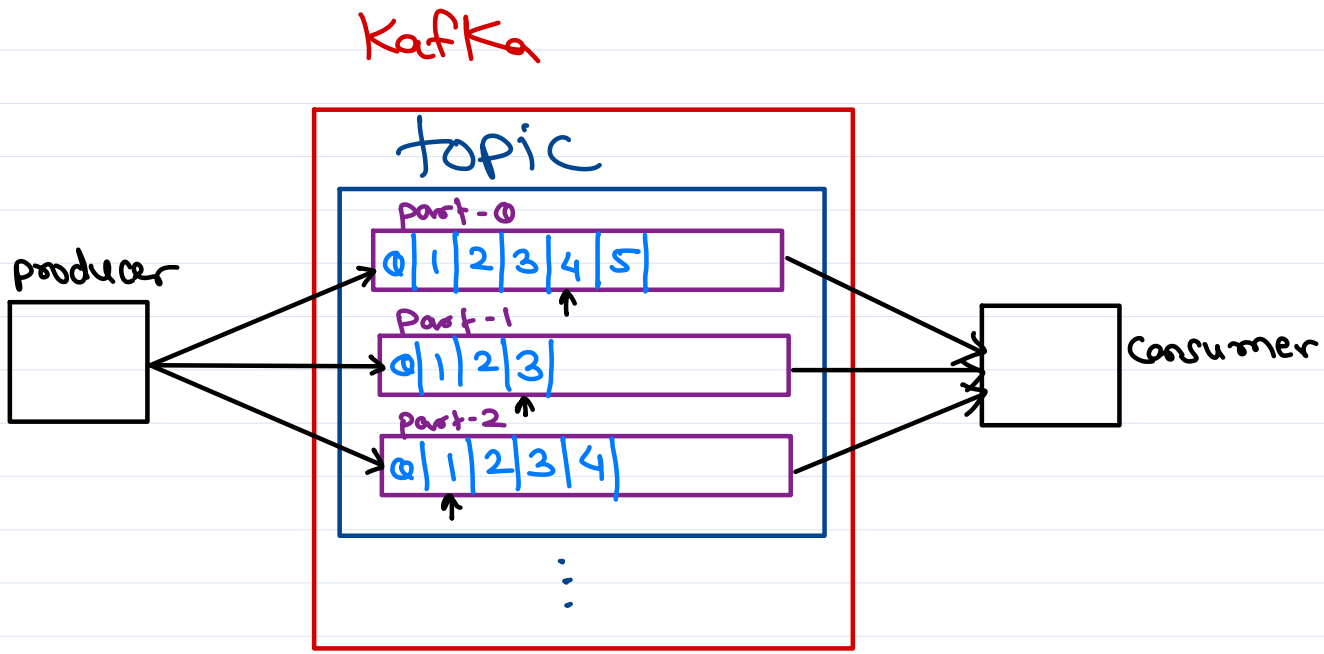




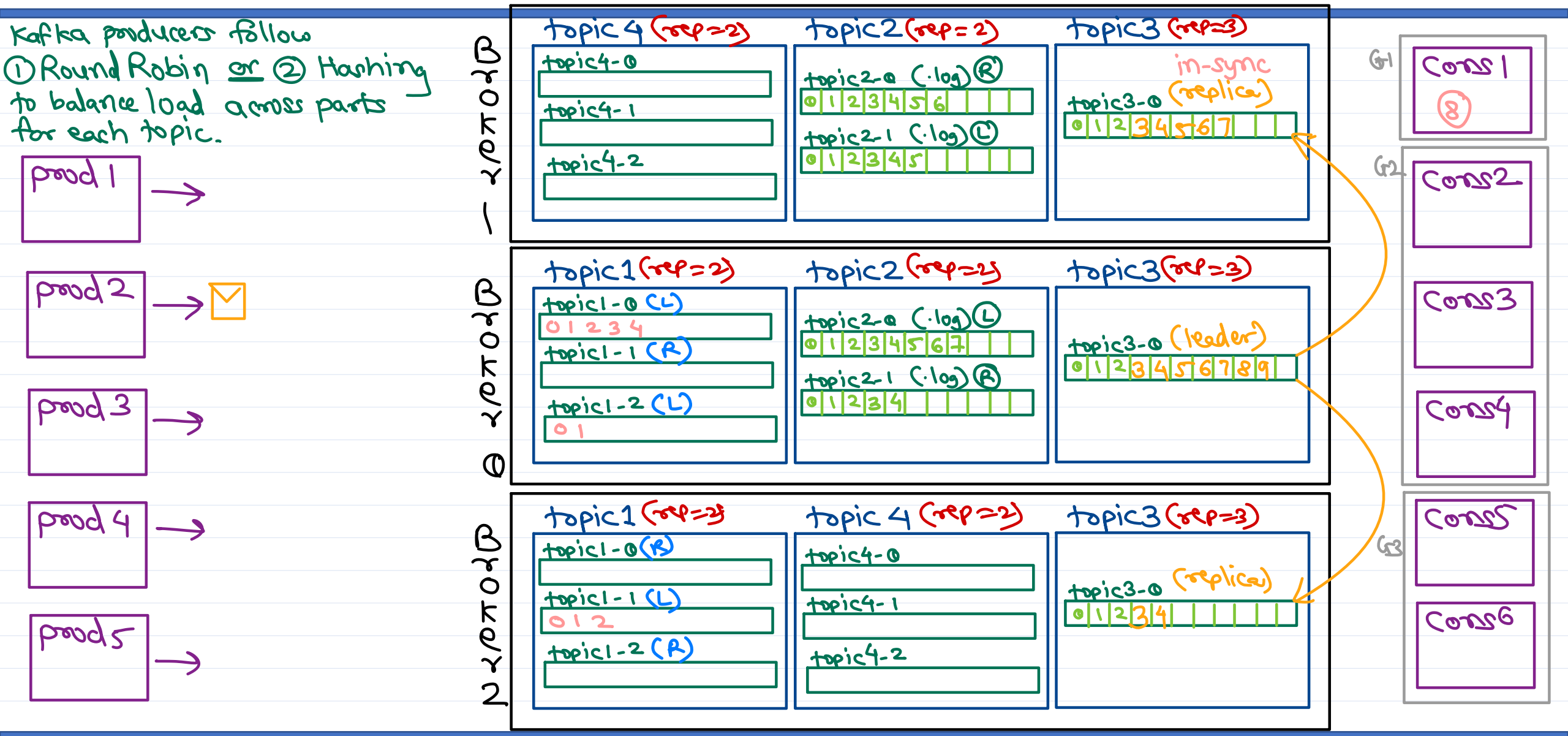


Kafka cluster and topics





Kafka Architecture



Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>