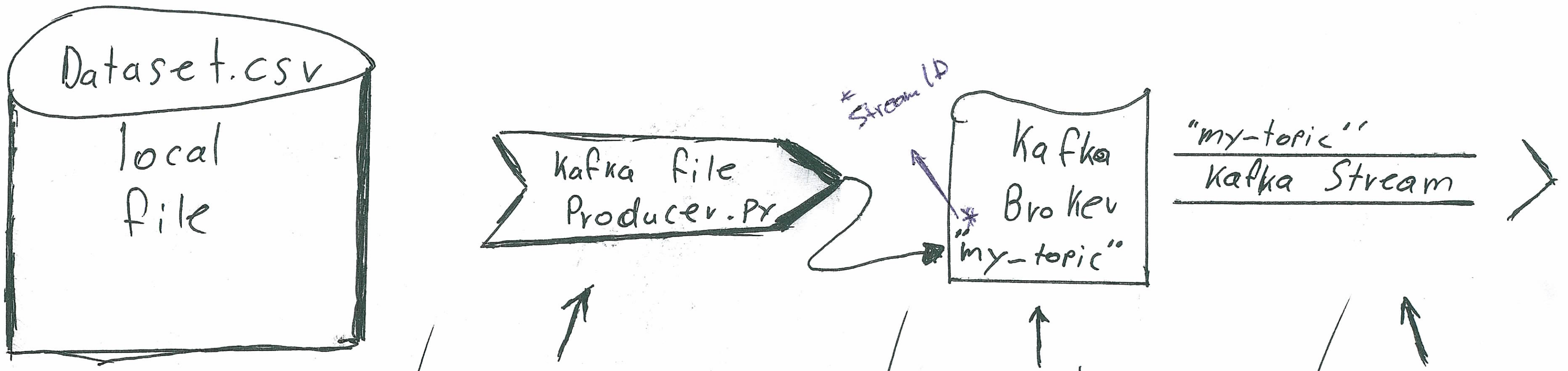
Stream Production



dataset resides in file in local FS

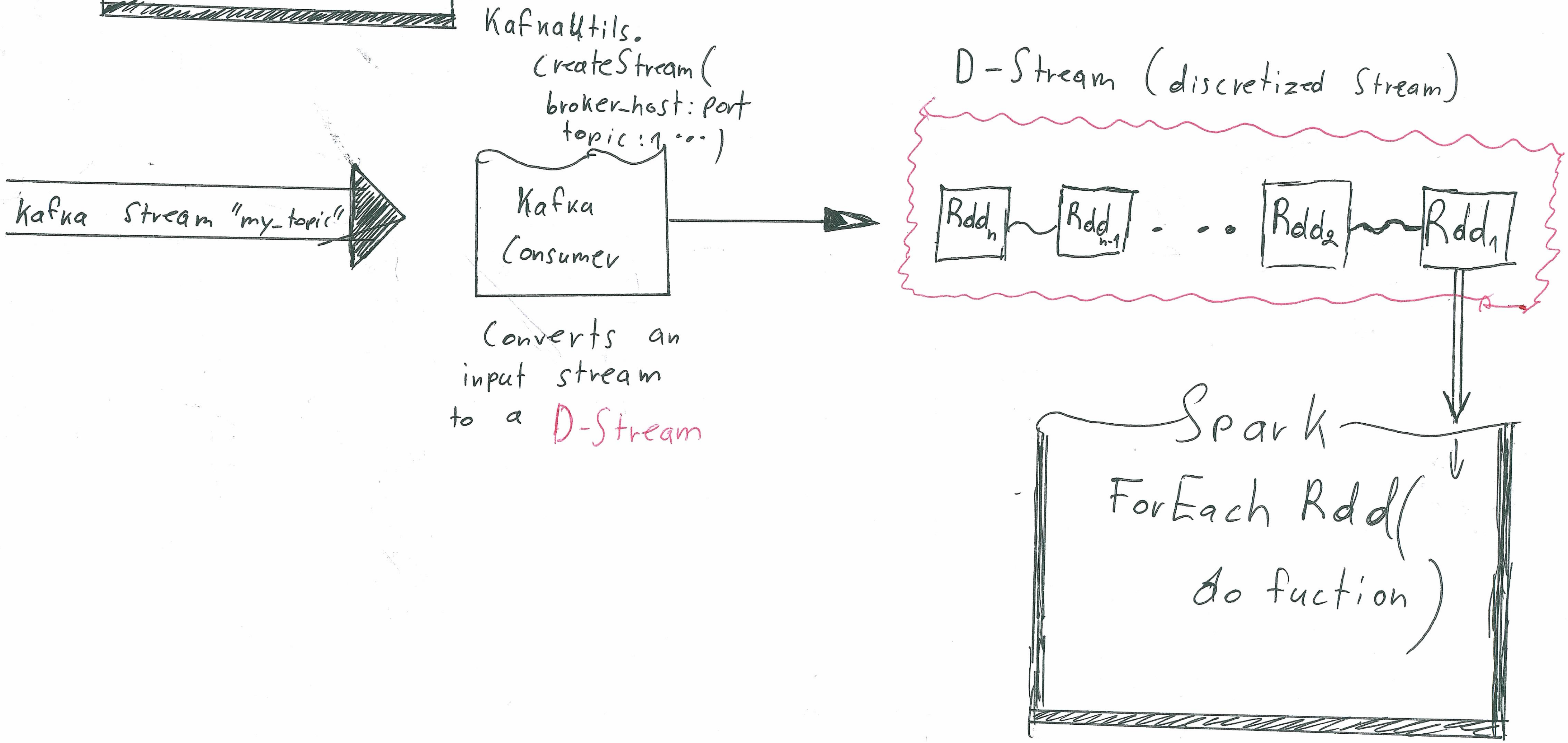
simple producer that reads the file and writes it under a topic to a Kafka broker

a host running Kafka (possibly one of many)

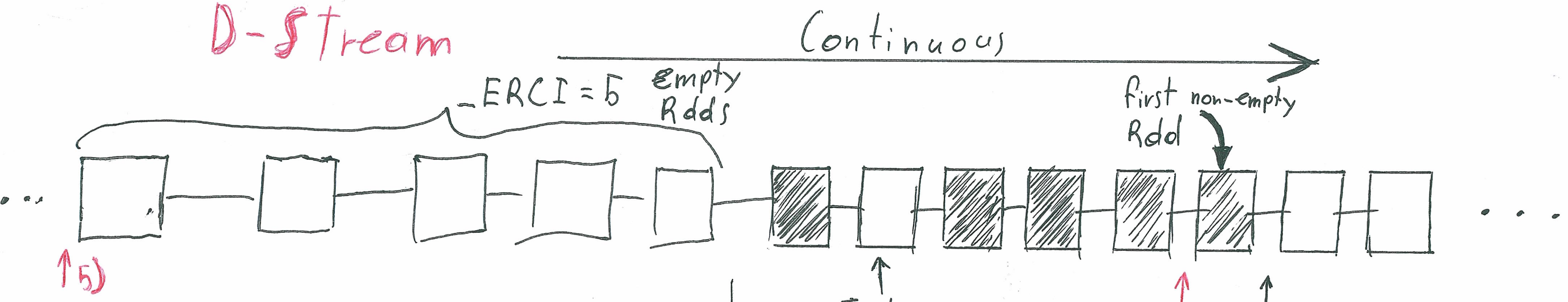
the resulting stream. A consumer can tap into a topic and consume (read) from this stream

Overview: Real simple: Some lines from a .csv file are written on stream under "my_topic"

Kafka Stream Consumption on Spark Streaming



Spark Streaming Custom Experiment



1) This is where we can realize that an experiment has been started

2) update time delta (based on prev_time and now)

3) add "now" time to rdd times and rdd.count to "records"

4) Printout

5) If the last _ERCI batches (Rdds) were empty, we deduce that the experiment has ended

Empty Rdd.: A single one, or (less than _ERCI) does not mean that the experiment has ended

1)

Start time = PRevious (empty) Rdd end time

rdd_times[0]

rdd_times[1]

end result

rdd_times: [0 , t_1 , t_2 , ... t_n]

(rdd_)records[0 , $Count_1$, C_2 , .. C_n]