Assumptions :- we are using kafka as our streamdata source

Event de duplication from streaming data sources(Kafka) can be done in a lot of different ways which are as follows.

1) We can use off the shelf solutions like ksql and using simple sql we can remove all the duplicates.
2) Storing message keys in a in-memory cache(redis), or fast key value look up stores like HBase running on top of S3 and removing the key from a pre-defined timeout.

But Since we need to store the count of duplicates it is better to go with approach 2.
As far as choosing the storage solution is concerned we should run tests to figure out which one fits the SLA, performance and cost parameters.

Using redis can give a very optimal performance but will be costly. With HBase running on S3 cost of infra will be less but latency will be higher than redis.