Within Session Personalization

Agenda

1. Kafka

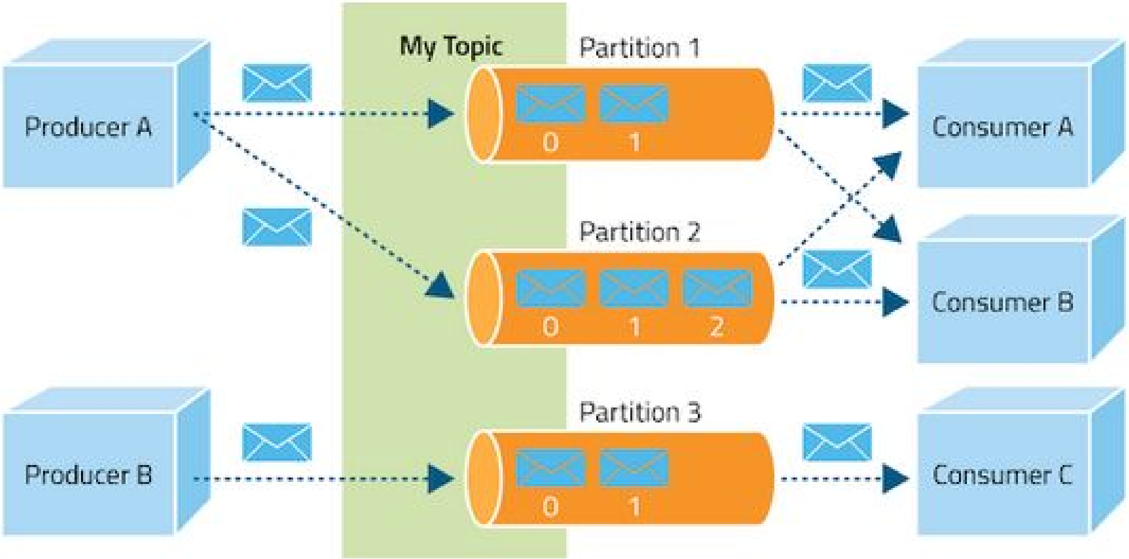
2. Kafka Streams

3. Sequence Based Model

4. Validation

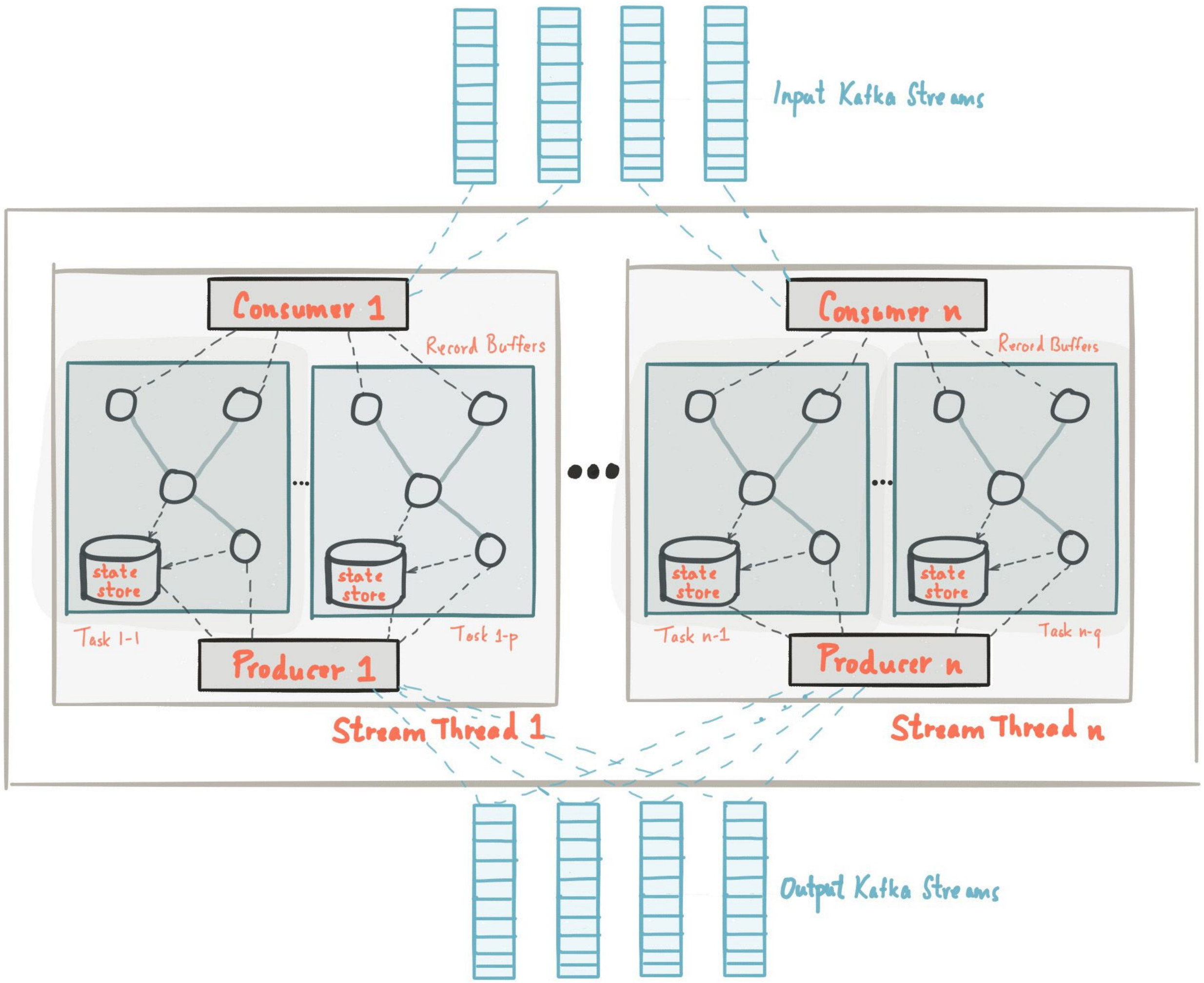
5. Status + plans

Kafka



- Kafka: distributed message queue; cluster of brokers (Etsy: 24 brokers)
- Topics: each topic like a physical log (Etsy: beacon-main)
- Ordered partitions; partitioned by key (Etsy: 144 partitions)
- Replication of partitions: leaders & followers (Etsy: 3 replicas)
- Producers (Etsy: franz)
- Consumer groups: each partition assigned only to one consumer in group (Etsy: orloj, hadoop)

Kafka Streams

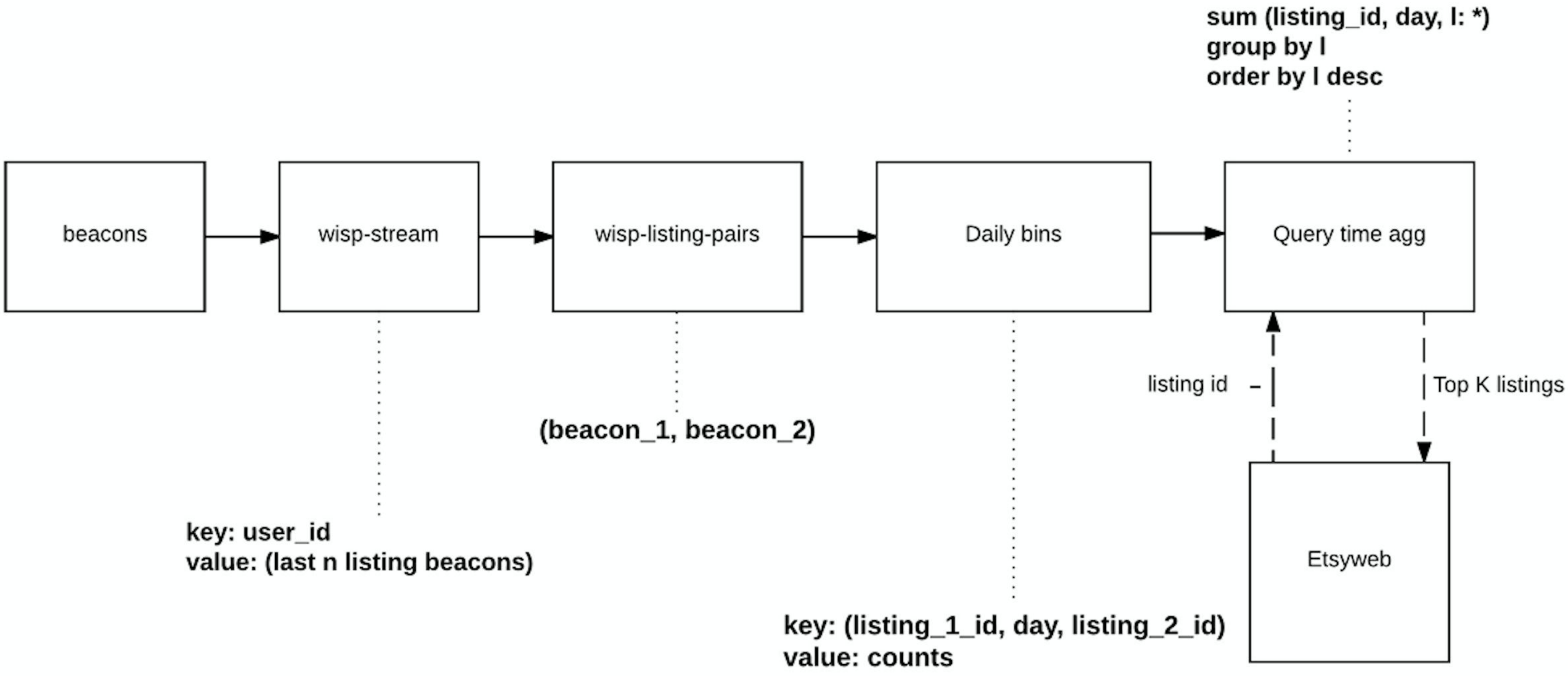


2 instances of the same Kafka Streams application that consume data from different partitions of the same Kafka topic.

Kafka Streams

- Kstreams & Ktables duality: Kstreams are the changelog of the Ktable
- Local storage in a key value store (rocksdb); checkpointed periodically to cluster
- dsl like scalding that allows transforming the data
- Eg:- can do joins on incoming streams
- Notion of time: processing time or event time
- Multiple instances of kafka applications; each instance can have multiple threads; each thread handles one partition

Sequence-based model



sum (listing_id, day, l: *)
group by l
order by l desc

(beacon_1, beacon_2)

key: user_id
value: (last n listing beacons)

key: (listing_1_id, day, listing_2_id)
value: counts

Sequence-based model

- Simple statistical markov model
- Count all listings historically clicked immediately after the listing in question
- Offer top k listings by frequency

Validation

- Training on last 60 days
- Testing on latest date
- A positive example is if the 2nd listing in a pair is in the ground truth set; negative otherwise
- AUC for sequence based beats the currently used similar listings which is based on tfidf scores ranked by quality scores (user signals from search)

Status + Future Work

- Working Kafka streams part on VM
- Need to figure out:
  - Latency; need to start using prod data
  - Removing data older than 60 days
  - Bootstrapping since EventPipe Kafka cluster has only 7 days worth of data
- Data-eng working on Kafka Apps AR whose scope includes deployment, failover, maintenance, and monitoring of kafka apps.
- Logging/Graphite calls from apps