CLOUD
COMPUTING
APPLICATIONS

Intro to Stream Processing

Roy Campbell & Reza Farivar

Thanks to public domain slides Jiangjie (Becket) Qin

What is Stream Processing?

Imagine you are browsing:

If you see an advert on a page, there will be an AdViewEvent

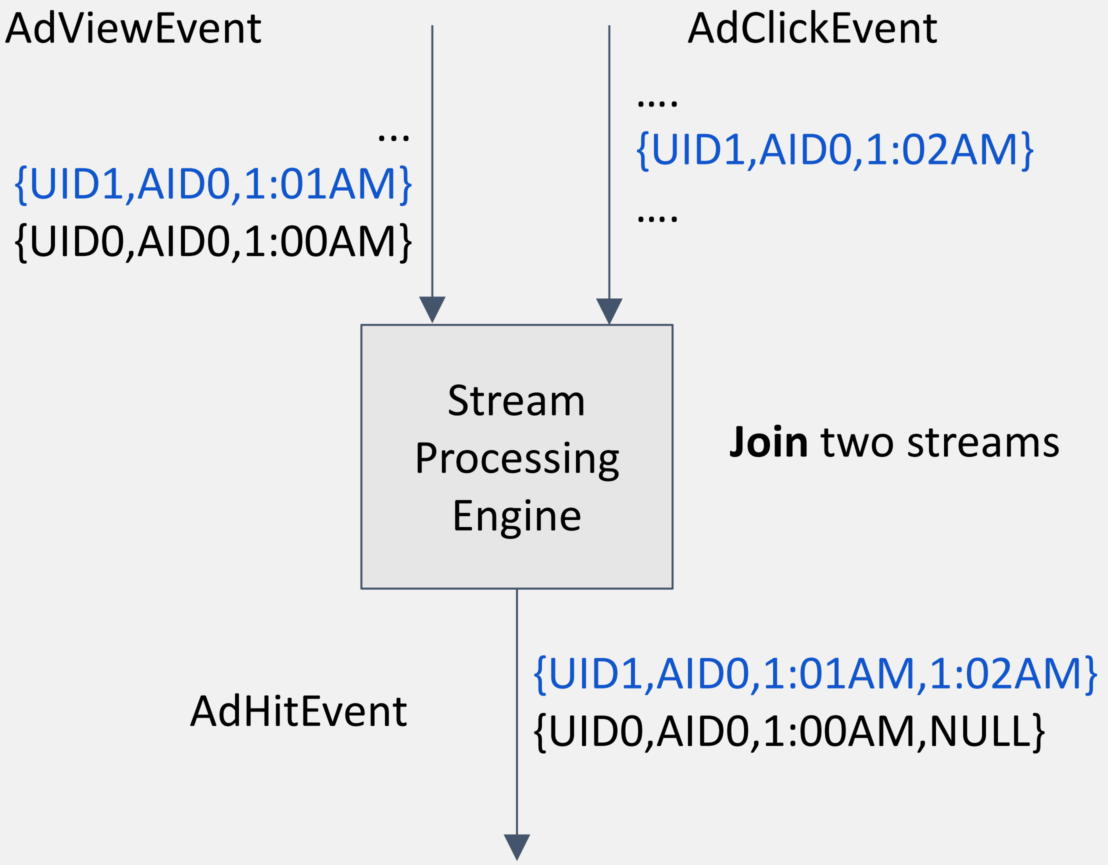
  - {UserId, AdId, Timestamp}

- If you clicked the ad, there will be another AdClickEvent
- {UserId, AdId, Timestamp}
- Now how can we know which is the most effective
  ad during last hour?

What is Stream Processing?

- Input – potentially infinite sequence of events
  - e.g. AdViewEvent, AdClickEvent
- Latency - near real-time
  - From milliseconds to minutes instead of hours to days
- Output - an infinite sequence of changes to the derived dataset
  - Another interim stream for further processing
  - The final result to store in the data store

What is Stream Processing?



AdViewEvent

...

{UID1,AID0,1:01AM}

{UID0,AID0,1:00AM}

AdClickEvent

....

{UID1,AID0,1:02AM}

....

Stream Processing Engine

Join two streams

AdHitEvent

{UID1,AID0,1:01AM,1:02AM}

{UID0,AID0,1:00AM,NULL}

The Requirement for Stream Processing

- Low latency
- Tolerate out of order and late arrival
- User friendly interface - streaming SQL
- Scalability
- Data safety and availability
- And others

Stream Processing

- What are the application requirements?
  - Scalable, fast, stateful stream processing
- What scale should we operate at?
  - Traffic Volume: 1.4 Trillion events/day
  - Intermediate State Size: multi TB / colo (*)
- Why is it expensive to run stream processing at scale?
  - Intermediate data set needs to be stored to allow low latency processing
  - Large volume of data needs to be pulled and pushed via network