

# Use of shared handles for Cache reuse across DoFn's in Python

By Amruta Deshmukh



BEAM  
SUMMIT

Austin, 2022



# Agenda

- Who am I?
- Strivr's stream processing platform
- Use Case
  - Initial approaches
  - Solution that worked
- Q & A

# Amruta Deshmukh

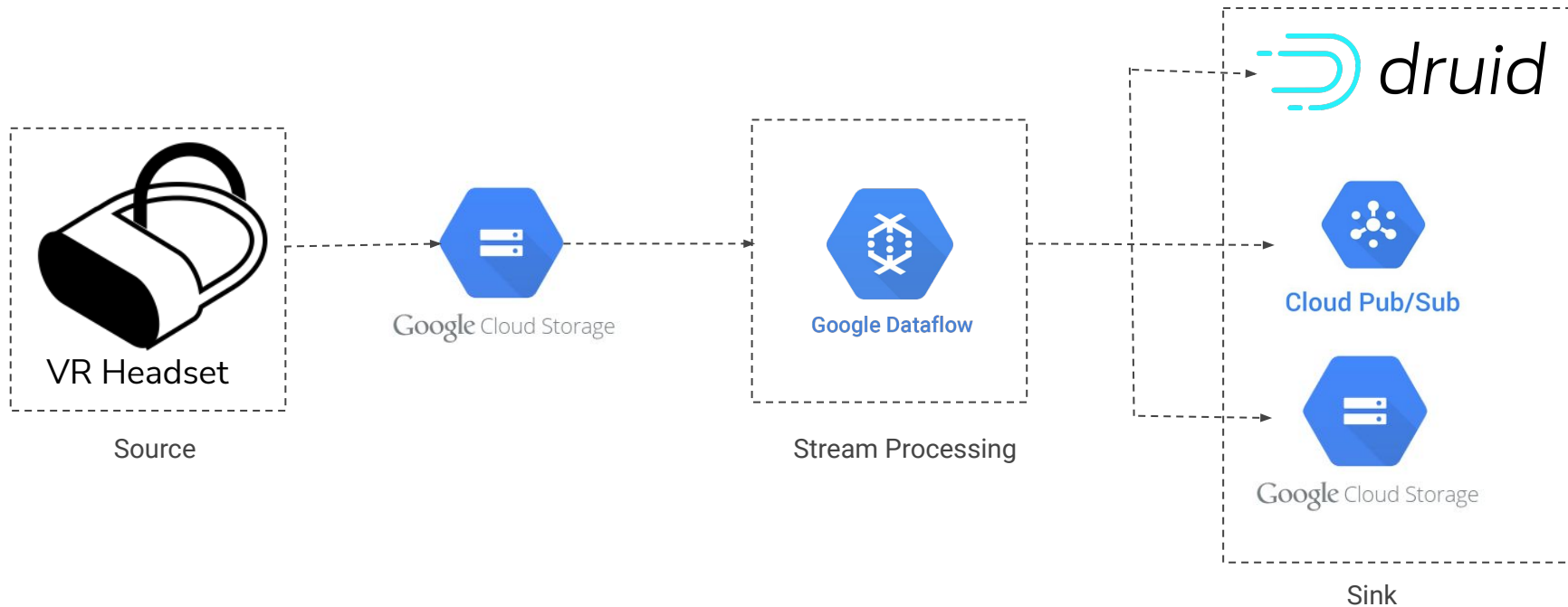


- Senior s/w Eng
- Member of the Data Platform team
- 2 years at Strivr



- Immersive Learning platform
- End-to-end solution for immersive learning at scale

# Strivr - Stream Processing platform



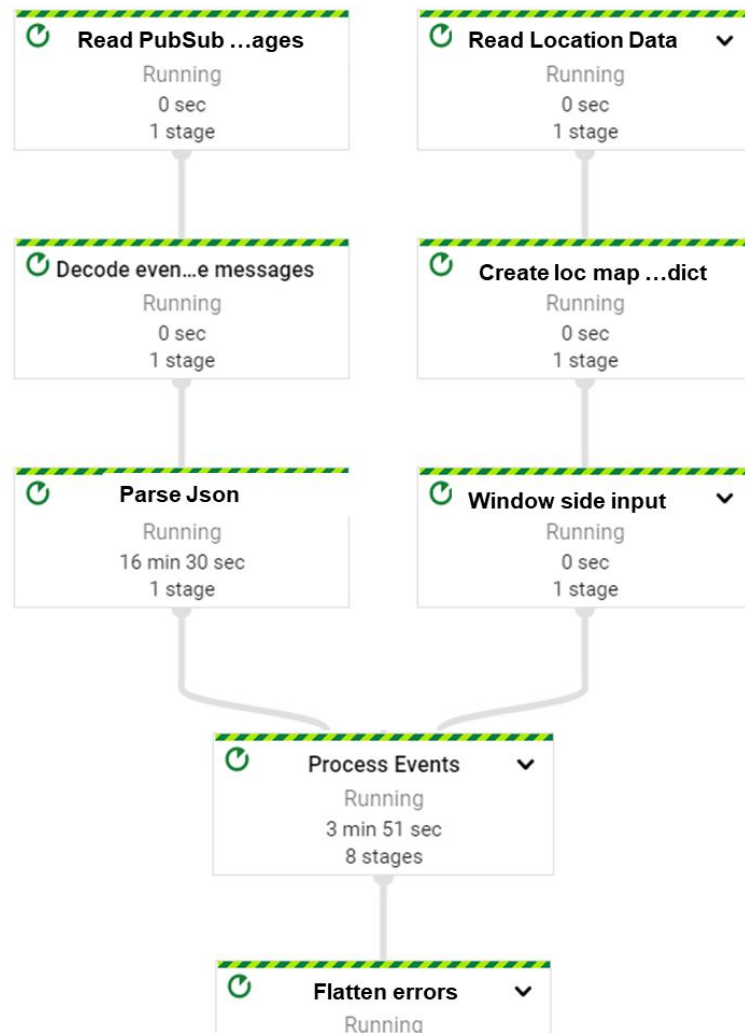


# Problem Statement

- Processing events received from the headsets
- Device metadata available in our Database backend
- Enrich events with the metadata (location)
- Metadata size less than 1 MB

# Side Inputs - Approach

- Read Location metadata from file in GCS
- Embed each event with location metadata
- Refreshing the side input every 1 hour





# Side Inputs - Issues

- Upon deployment we started seeing issues
- Errors messages
  - "Error message from worker: generic::data\_loss: SDK claims to have processed 90959 but should have processed 90077 elements"
- Dataflow pipeline autoscaled workers
- Worsening data freshness
- Pipeline not successfully draining



# Stateful Dofn - Approach

- Group events by device id
- Using stateful Dofn to read location metadata from Datastore
- Append the metadata to the events
- Refresh the state every 1 hour





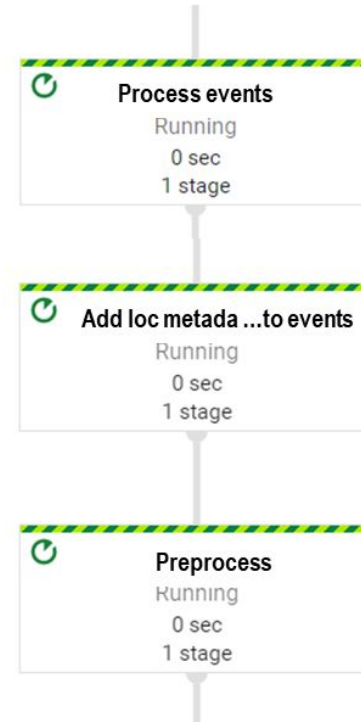
# Stateful Dofn - Issues

- Noticed major bottle neck at the stateful dofn step
- Increased data freshness from that step
- Rewindowing to global windows and fixed windows could possible issues downstream



# Shared cache - One that worked!

- Read the location data from a GCS file
- Load the file into a shared cache object as a dict
- Refresh cache object after a fixed time interval (1 hour)
- Given the timestamp of the current event gets either the cached object or refreshes it from GCS file



# Sample code



- Here is the [repo](#) with the sample code

# Questions?




BEAM  
SUMMIT

Austin, 2022

# Thank You!

Let's connect!

: <https://www.linkedin.com/in/adeshmuk/>

: <https://github.com/amruta-d>



BEAM  
SUMMIT

Austin, 2022