Hi Nathaniel,

My friend and I are working on tooling that makes it easier for companies to work with their user data responsibly. I'm writing to you to ask if you would find our products useful.

Background. My friend, Rohan, and I studied CS at Princeton and went to work for Google after graduation. We both worked on teams that use user data to make smarter products: I worked on video features in Search (you can see my baby, the Sports Key Moments project, in this blog post), and Rohan worked on accessibility features for Assistant.

Problem. We spent a lot of effort ensuring our use of user data was compliant with policies like GDPR. The high level problem we faced was that we needed to ensure all the data we wanted to use in our work would be deleted within a certain timeframe, or, as soon as possible, if a user requested it. We accomplished this through a hacky mix of spreadsheets and scripts, and it was a huge time-sink.

Solution. Our idea for an improved solution is: create a layer on top of each user data store, which can programmatically enforce different policies. More specifically, we're imagining a situation like this:

- Hinge Health has an S3 bucket containing tables of user session logs.
- Hinge Health promises users it will delete non-essential data collected from them after 30 days.
- Data scientists routinely query the tables, say, with MemSQL. Sometimes they need to transform the data and write derivative tables.
- **Status quo:** nobody except each data scientist knows where they wrote the data and what they do with it. The data scientist is responsible for cleaning up derivative tables.
- **New world:** a MemSQL plugin will record where any derivative tables are written. These tables will not be copyable via normal filesystem commands. A daemon will take care of deleting all derivative tables after 30 days.

The main question on our minds is: would eng orgs dealing with user data find this useful? If you have any thoughts about this, we would love to hear.

Thank you,
Stefan and Rohan