Q&A from Town Hall Meeting 03/20

Q: Why do DataHub?

A: DataHub is a self-service data portal which provides search and discovery capabilities (and more!) over the data assets of an organization. This tool can help improve productivity of data scientists, data analysts, engineers of organizations dealing with massive amounts of data. Also, the regulatory environment (GDPR, CCPA etc) requires a company to know what data it has, who is using it and how long it will be retained - DataHub provides a solution to these challenges by gathering metadata across a distributed data ecosystem and surfacing it as a data catalog thereby easing the burden of data privacy/compliance.

Q: What are DataHub's current challenges?

A: The team is working hard on externalizing (making generic) and open-sourcing some of the LinkedIn specific concepts and pages that we've built over the years, including compliance, lineage etc. We're also working hard to make the metadata platform and DataHub UI easy to use and extend. Our roadmap should give you a good sense of where we are headed.

Q: How does DataHub integrate with Data Sentinel?

A: Data Sentinel runs on datasets / data samples either synchronously when data is written or asynchronously after data has been written. The results of these data quality checks are stored in Data Sentinel's database and exposed using an API. Internally at LinkedIn, DataHub calls / will call the Data Sentinel API to surface the data quality results on the Dataset Health page. When we open source the code on the DataHub side, we will make it pluggable so that other deployments can add their data quality provider. We might even open source a data quality metadata model for teams that want to store the data quality results integrated directly in the DataHub GMS instance.

The Data Sentinel team does not have any short-term plans to open source it yet, but we will keep the community informed if that changes anytime!

Q: How granular is the data lineage that you track at LinkedIn? Do you also try to track field-based lineage? E.g. for one particular column that enters a kafka stream, would you be able to trace down all the tables that later are related to this column? And do you also track transformations to this column while it travels the DAG (e.g. filter for certain values, group by another dimension)?

A: Currently we only track coarse grained lineage, however we do plan to extend it to field-level lineage as covered in the roadmap as well.

We plan on using Apache Calcite to parse and represent fine-grained lineage. LinkedIn has open sourced it's query translation capability Coral. It was built to translate from one declarative language to another e.g. going from Pig to Spark or Spark to Samza SQL, etc. Coral will allow us to build a lot of the translators that go from commonly used languages or frameworks to write code to an intermediate representation using Calcite. We plan to use the intermediate Calcite AST to generate fine grained lineage for various kinds of jobs and flows that people are writing. Calcite gives you the query AST - you can use it as your internal data structure to transfer metadata properties over or simply build links between input columns and output columns.

Q: Would you recommend Datahub rather than available commercial solutions?

A: Common problems with commercial solutions can be summarized as:

1. Lacks direct access to source code: Any feature gaps can only be closed by external parties, which can be both time consuming and expensive.
2. Dependency on larger proprietary systems or environments, e.g. AWS, Azure, Cloudera etc., making it infeasible to adopt if it doesn't fit your environment.
3. Expensive to acquire and operate.

Datahub can be right for you if you want an open source unbundled solution (front-end application completely decoupled from a "battle-tested" metadata store), that you are free to modify, extend and integrate with your data ecosystem. In our experience at LinkedIn and talking to other companies in a similar situation, metadata always has a very company specific implementation and meaning. Commercial tools will typically drop-in and solve a few use-cases well out of the gate, but will need much more investment or will be impossible to extend for some specific kinds of metadata.

Q: Does Datahub have some data quality functionality?

A: See discussion on Data Sentinel above.

Q: How are you thinking about Job entities and their relationship to lineage?

A: We are very close to finishing up our coarse-grain lineage rollout at LinkedIn (ETA June 2020 inside LinkedIn). This introduces the concepts of Jobs and Flows and connects them with input and output datasets. While our first version of the metadata model refers to Azkaban Flows and Jobs, it can easily be extended by the community to support other open source schedulers like Airflow etc.