

Report

The point of this paper is that metadata is important. Since the data over the past 10 years has grown up by 40x, it has introduced the following 2 problems:

1. Productivity: How is the new build structure used to bring out effectiveness and productivity.
2. Compliance: How do organizations comply with increasing regulatory, and compliance demands and uphold the trust of their users?

The solution of the above-mentioned problem is metadata. What is metadata?

1. Describing set of data
 - A- Application context- details needed by us
 - B- Behavior- about the data
 - C- Change- how will it change overtime
2. The data being described
 - It includes everything like Data stores, Schemas, Dashboards, Streams, Processing and people.

It was observed that a lot of the time was being spent in data discovery while Lyft team was majorly focused on model development and productionalization.

Data discovery answers the questions like Can we trust the data, Can I reuse my existing work or not and other like who are users and does this data exist?

Amundsen is used for search for data within the organization. It has two pages- Landing page and Detail page. Landing page- is the entry point where we can go and enter our search data, it also presents popular options to browse through it. Results include description and last updated about the table. Detail page- after choosing result of choice, this page shows the detail about the description, owner, frequent users and general profile of the data. This page also has feedback widget where users can provide their feedback. Amundsen has a higher result rate and customer satisfaction score, yielding higher trust.

Compliance: Such a metadata powered solution can help an organization remain compliant as the data and its use-cases or service requests grow.

With large amounts of data, the success in using data to fullest lies not in data but in the metadata. Lyft has built a data discovery platform, Amundsen, which has worked well in improving the productivity of its data scientists by faster data discovery.

At the same time, there's a lot of value a metadata driven solution can provide in the space of compliance, in tracking personal data across the entire data infrastructure.