Lyft’s data discovery & metadata engine

The article is dedicated to the Amundsen project which goes on to discuss the project’s goal and its users. The scientists at Lyft developed an application that was built on a metadata engine to aid data scientists and researcher scientists which was known as Amundsen. Amundsen was inspired by search engines like Google. It was achieved by developing a search interface for data.

Two challenges associated with growth in the data volume are productivity and compliance which could be solved with the metadata. The article further discusses the productivity problem and how it can be solved using metadata.

At Lyft it was observed that the data science team spent significant part of their time on model development and productionalization while they wanted to dedicate more for data discovery. To solve this problem Lyft came up with Amundsen.

The application developed has an entry point which is a search box where one can enter a plain text to search for data. The results depend on the metadata that contains the table description and the date when it was last updated. The text matching algorithms make the search successful. The ranking is similar to the page rank.

A detail page is also displayed to the user. With specific colors and behavior.

The articles further discusses the following trade offs that had to be made while building the application.

1. Discovery vs Curation
2. Security vs Democratization

In the conclusion the article highlights the significance of metadata and points out that the success in using data lies in its metadata. And Amundsen, a data discovery platform has been a successful project at Lyft. At the same time, metadata driven solutions offer valuable capabilities for tracking personal data across the entire data infrastructure, in order to comply with the law. This area is likely to see much investment in the future.