Lyft's Data Discovery and Metadata Engine

The Amundsen project, Lyft's data discovery and metadata engine, its objectives, and users are explained in this article. Amondsen is a data discovery application developed on top of a metadata engine with the goal of increasing the productivity of Lyft's data scientists and research scientists. To clarify the issue, the author claims that tremendous rise in data quantities has resulted in two major challenges: First, productivity – whether it's constructing a new model, instrumenting a new variable, or doing ad hoc analysis, how can we utilize this data most productively and effectively? Second, compliance - while gathering data on a company's users, how do enterprises meet rising regulatory and compliance standards while maintaining their customers' trust?

And as per the author, the key to resolving the aforementioned issues is not in the data itself, but in the metadata. Metadata is, at its heart, a collection of data that describes and provides information about other data. Metadata consists of two parts: a (typically smaller) collection of data that describes another (usually bigger) group of data. Three major forms of metadata are used to describe a set of data: application context, behaviour, and change (ABC of metadata). Capturing these three types of metadata and using them to drive apps is critical for many future applications. The ABCs above describe any data within an organization. Data Stores, Dashboards/Reports, Events/Schemas, Streams, Processing, and People are all examples of this. This precise metadata may be utilized to increase data consumers' productivity by putting pertinent metadata at their fingertips.

Lyft discovered that, while they intended the majority of their time spent on model building (aka prototyping) and productionalization, a large portion of their time was spent on data discovery. Amundsen was heavily influenced by search engines like as Google — in fact, within the business, it is frequently referred to as "Search for data. “After explaining the trade-offs in this project between discovery and curation, as well as security and democratization, the author concludes that with enormous volumes of data, the success in fully using data resides not in the data itself, but in the metadata. Amundsen, a data discovery platform developed by Lyft, has shown to be extremely effective in increasing the productivity of its data scientists through quicker data discovery. At the same time, a metadata-driven solution may give a lot of value in the area of compliance, tracking personal data across the whole data infrastructure. In the future, we should expect a lot more investment in that sector.