

Case Study



Lyft

Lyft Speeds Up Data Discovery with Tool Using Neo4j

INDUSTRY

Transportation

USE CASE

Data Discovery / Knowledge Graph

GOAL

Enable Lyft employees to quickly find the right data to analyze

CHALLENGE

Massive growth and data accumulation resulted in difficulties finding data

SOLUTION

Create a tool that accelerates data discovery using Neo4j

RESULTS

- Lyft data scientists are 30% more productive
- Users give Amundsen a customer satisfaction rating of 8.5/10

At Lyft, finding the right data to analyze was difficult, and with growth in data volumes and team members, the problem was getting worse. Lyft built Amundsen to streamline data discovery.

The Company

Based in San Francisco, [Lyft](#) is an app-based gig economy company that provides more than 50 million rides a month to customers through its network of contract drivers. Riders use the Lyft app, which generates data with every ride Lyft provides. Lyft is Uber's main competitor, and operates in 640 cities in the U.S., as well as nine cities in Canada.

The Challenge

Data is at the heart of every decision at Lyft. Once decisions are made, their impact is evaluated using data.

Given the vital role of data and analytics across the company, the speed with which users can find data, understand it, analyze it and gain insights is critical.

Data discovery – finding the right data and understanding it – was slow and inefficient. Tables might have similar names, like `driver_rides_completed` and `rides_driver_total.lifetime_completed`. Users asked coworkers for help, reached out on Slack channels or looked at Github to see how a table was generated. They often pulled the first 100 rows to get a feel for the contents.

Lyft's growth exacerbated the challenge of data discovery. Lyft already had about 10 petabytes in thousands of tables across a variety of different data stores according to Tamika Tannis, a Lyft software engineer. Growth meant even more data generated by the mobile app and other services. As new talent was hired, the number of users doing data discovery also grew.

Lyft needed a better way to support data discovery for everyone in the company. To quantify the problem and get a baseline, Tannis's team looked at the impact on data scientists and found that data discovery consumed about a third of their time.

The Solution

Lyft engineers decided to build a tool to simplify data discovery. Their first target audience would be the most frequent users of data: analysts and data scientists.

Named [Amundsen](#), the tool would offer three complementary ways to do data discovery: search-based, lineage-based and network-based.

An effective search was a top priority, ranking results by popularity and relevance. Lineage-based discovery traces connections among datasets. Network-based data discovery connects data with people, particularly valuable for new team members.

Case Study



“Amundsen has been really successful at Lyft. We currently have engineers, product managers and even customer service folks using Amundsen to find what they need.”

—*Tamika Tannis,*
Software Engineer, Lyft

“You might want to see what data resources your manager or your coworkers are using so you can use trusted data resources that everyone else is already using for similar purposes,” said Tannis.

Amundsen uses a microservice architecture. The Databuilder service ingests data into the search service, which is backed by Elasticsearch, and the metadata service, which is run by the [Neo4j graph database](#). Elasticsearch powers the search by providing relevance based on search terms, the user’s position in the company and the popularity of the tables. All of those connections are first made in Neo4j.

Lyft chose Neo4j because it captures the shape of their data ecosystem, which is naturally expressed as a graph. The flexibility of Neo4j is very beneficial when it comes to iterating quickly on new features.

“When we have a new use case and a new piece of metadata to represent, we just have to create a new node and create that relationship,” said Tannis.

At Lyft, Neo4j is an important component of Amundsen’s architecture; it serves as the source of truth for editable metadata. Neo4j also provides a foundation for new projects like compliance and data quality. “The future, as I see it, is that we’ve got a full-fledged metadata repository on which we’re building many applications,” said Mark Grover, a product manager at Lyft.

The Results

Amundsen adoption is high. “We have 90% penetration among data scientists, which was one of our target personas,” said Tannis. “What this means is that 90% of data scientists are using Amundsen to do their jobs on a weekly basis. We also found that this tool has increased productivity for our entire data science organization by around 30%,” noted Tannis.

Even before the team iterated Amundsen for additional personas, others started using it. “Amundsen has been really successful at Lyft. We currently have engineers, product managers and even customer service folks using Amundsen to find what they need,” said Tannis.

Users provide feedback right in the app. This gives users a voice, from rating the tool to requesting new features. “If something doesn’t look right, they can immediately file a bug report,” explained Tannis. Amundsen gets high marks from Lyft employees. Its customer satisfaction (CSAT) score is 8.5 out of 10.

To make the impact of the tool even broader, Lyft chose to open source Amundsen. An active community is growing up around it, and more than 10 companies now contribute to the project.

Neo4j is the leader in graph database technology. As the world’s most widely deployed graph database, we help global brands – including [Comcast](#), [NASA](#), [UBS](#), and [Volvo Cars](#) – to reveal and predict how people, processes and systems are interrelated.

Using this relationships-first approach, applications built with Neo4j tackle connected data challenges such as [analytics and artificial intelligence](#), [fraud detection](#), [real-time recommendations](#), and [knowledge graphs](#). Find out more at [neo4j.com](#).

Questions about Neo4j?

Contact us across the globe:
info@neo4j.com
neo4j.com/contact-us