SUCCESS STORY

INDUSTRY: Ride Sharing

LOCATION: San Francisco

PREVIOUS SOLUTION: None

KEY BENEFITS:

Deliver data lineage transparency and literacy
Enable cost-effective, confident data migrations
Reduce risk posed by corrupt or inaccurate data
Optimize compute resource utilization, savings millions
Improve productivity at every level
Ensure data accuracy

CompilerWorks®

www.compilerworks.com | info@compilerworks.com

> CompilerWorks Lineage has helped close the gap that existed in our lineage knowledge. It has helped us do impact analysis for our data migrations, upstream and down."
>
> Liuyin Cheng, Lyft Data Engineer

COMPANY OVERVIEW

Lyft was founded in 2012 and is one of the largest transportation networks in the United States and Canada. As the world shifts away from car ownership to transportation-as-a-service, Lyft is at the forefront of this massive societal change. Our transportation network brings together rideshare, bikes, scooters, car rentals and transit all in one app. We are singularly driven by our mission: to improve people's lives with the world's best transportation.

CHALLENGES

Founded in 2012 and based in San Francisco, Lyft provides millions of rides daily as the world shifts away from car ownership to transportation-as-a-service.

Lyft is technology driven and seeks to continuously improve the Lyft App at the core of its services. From the beginning, Lyft has relied on a cloud-based infrastructure, particularly Amazon Web Services (AWS), including Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).

Initially, Lyft was dependent on Amazon's Redshift data warehouse and the Kinesis message bus, but ran into scalability issues due to the tight coupling of compute and storage. The firm needed to provide fast, flexible access to petabytes of both structured and unstructured data to power a vast number of data marts, BI tools, CMS apps, and a growing force of data scientists and engineers. Lyft elected to migrate from Redshift to Apache Hive, yet still host in the AWS cloud. Today, Lyft has more than 100,000 tables in Hive and a few thousand left in Redshift.

"We were continually moving data from Table A to Table B via ETL," notes Lyft data engineer Liuyin Cheng. "Before we discovered CompilerWorks Lineage, we had little insight into our true data lineage. We did not have a good idea of the impact of changes in our data flow and access. We needed to be able to visualize our information flow, track data errors, and conduct impact analyses of data pipeline changes."

THE SOLUTION

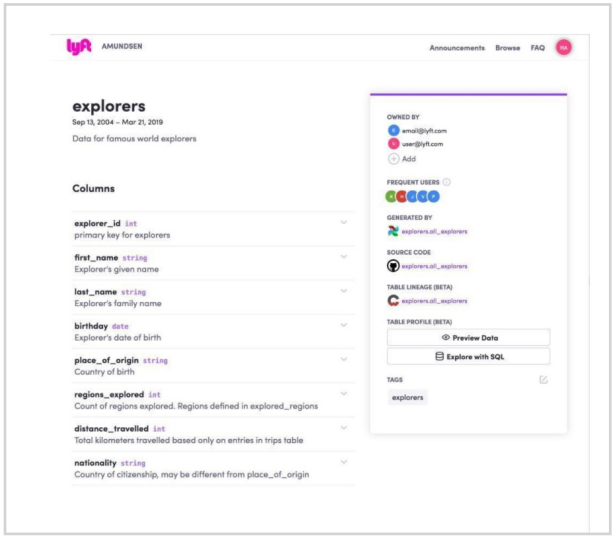
At Lyft, data are constantly increasing, including SQL tables and views in Redshift, Presto, Hive, PostgreSQL, as well as dashboards in BI tools like Mode, Superset and Tableau. To provide faster, better access to targeted data, Lyft developed its own data discovery tool—Amundsen (after the Norwegian explorer, Roald Amundsen).

Lyft employs CompilerWorks Lineage to enable Amundsen users to trace the lineage of data based on logs, code, BI reports, ETL pipelines and other sources. The system automatically creates a unified model or lineage fabric that shows what data are used, by whom, for what and how the data were processed.

THE SOLUTION (CONT.)

Data migration via ETL pipelines is a continuous and critical activity for Lyft data engineers. ETL pipelines are managed using Apache Airflow and data lineage assessment via CompilerWorks Lineage is an important step in the process. The firm also uses the system to ensure data quality is maintained, pipeline processing costs are optimized, usage is carefully tracked, and new hires are provided with data transparency at all times.

"Before, we had to try to manually trace the lineage of data the best we could," Cheng notes. "With CompilerWorks Lineage, when we have a need to know the lineage of a data pipeline, we have a powerful tool that tells us all we need to know. It's saving us a lot of time and improving our productivity. Understanding the scope of the data and who will be impacted by the proposed changes is a big part of the project, and CompilerWorks helps us do that."



Lyft's Amundsen data discovery application directs users to Compiler-Works Lineage for table lineage discovery and relationship correlation.

THE RESULTS

Deployed at Lyft in 2017, CompilerWorks Lineage is an indispensable tool that has helped the firm's data scientists, engineers and business users achieve more timely information delivery, greater accuracy, improved reliability, compute/storage cost control, and efficient accounting of data sources and value. Let's take a look at some use cases for CompilerWorks Lineage at Lyft:

"Before we discovered CompilerWorks Lineage, we had little insight into our true data lineage. We did not have a good idea of the impact of changes in our data flow and access. We needed to be able to visualize our information flow, track data errors, and conduct impact analyses of data pipeline changes."

Liuyin Cheng, Lyft Data Engineer

PIPELINE EXPLORATION

Details: Lyft uses Amundsen to explore its data catalog. However, Amundsen is static in nature—it shows the tables that are available in the database, but does not provide pipeline context (e.g., how a table is populated from upstream tables, how a table contributes to the population of downstream tables, etc.). In order to explore this context, Lyft has configured Amundsen so that each table detail page lists a URL linking to that table in CompilerWorks Lineage. Without this context, it is difficult to navigate a large and evolving catalog of tables and pipelines and derive any actionable meaning—it is akin to a glossary of terms without any definitions.

Business impact: CompilerWorks Lineage enables Lyft to visualize and explore the full picture of how data flows through the organization in an accurate, automated and continuously-updated fashion. This is the cornerstone that enables all of the additional use cases listed below.

DATA QUALITY

Details: Lyft's Data Engineering team receives data from upstream sources—mainly event data generated by Lyft's apps and servers. This upstream data can either have discrepancies in and of itself or cause downstream discrepancies. Lyft uses CompilerWorks Lineage to track the discrepancy from its discovery back to its source, and ultimately to fix the code that caused it.

Business impact: Lyft has reduced the risk posed by the propagation of discrepancies. "Essentially, the impact is equal to the cost of making a bad business decision, which could be huge," notes Cheng.

PIPELINE MIGRATION

Details: As Lyft's business grows and changes, its data pipelines must adapt. Often, this means adapting both the business logic and the underlying data processing technology that the logic is executed on. At Lyft's scale, this can result in migrations on the order of 200+ tables. These complex migrations typically involve three phases: planning, code conversion and monitoring. CompilerWorks Lineage can accelerate all three of these.

For planning, Lyft uses lineage to quickly and economically identify all the tables that must be migrated, the order in which they must be migrated, and which downstream teams will be affected and should be alerted. For monitoring, Lyft runs pre- and post-migration pipelines in parallel for a period of time and uses the CompilerWorks Lineage API to monitor adoption of the post-migration pipelines. Analysts run a job that extracts and filters recent usage data for both the pre- and post-migration pipelines, dumps the data to tables and compares them. This strategy allows Lyft to hedge the risk intrinsic to pipeline migrations while also ensuring that the migration is successfully adopted and the pre-migration pipelines are eventually decommissioned.

Business impact: CompilerWorks Lineage allows Lyft to save time, reduce cost and hedge risk when it comes to large-scale data pipeline migrations.

COST CONTROL

Details: Lyft uses CompilerWorks Lineage to identify pipelines that continue to consume compute resources despite the fact that their outputs are no longer referenced by downstream users. Once identified, these pipelines can be decommissioned and the compute resources can be reclaimed for other tasks, or decommissioned in turn. Lyft monitors this activity using the CompilerWorks Lineage API.

Business impact: The business impact here depends on how many unused workloads Lyft is able to identify: more unused workloads identified = more reclaimed/decommissioned servers = more cost savings. As a result, it is estimated that Lyft is saving millions of dollars per year as a result of this compute resource optimization.

USAGE TRACKING AND REPORTING

Details: Lyft is able to track the actual daily usage of delivered data marts and reports via CompilerWorks Lineage. The firm is able to decide when a depreciated report can be retired, and to track the user acceptance of the newly released updates. Many data shops operate on a "if we build it, they will come" basis in which they are blind to the usage patterns of their products. While the usage reporting built into specific reporting tools (e.g., Mode, Tableau, etc.) can offer a pinpoint view of certain layers in the data stack, only CompilerWorks Lineage offers a one-stop shop for usage analysis across all reporting tools.

"It definitely saves us money on the computational side to be able to use CompilerWorks Lineage to monitor table usage and identify those that aren't being used anymore and shut them down," Cheng says.

Business impact: CompilerWorks Lineage allows Lyft to use downstream usage as a proxy of ROI on their data engineering investments. It ensures the data engineering team is building pipelines that provide the most value to the business.

ONBOARDING NEWHIRES

Details: CompilerWorks Lineage is linked in the Lyft onboarding overview, and is used to get data engineers and scientists up to speed.

Business impact: Data engineers and scientists are expensive. Getting them acquainted with the landscape of data pipelines can ordinarily be an arduous and long process. Saving time in this process is critical and cost savings here is considerable with a staff of ~200 data engineers and scientists.

Lower costs, increase performance and streamline management for all your data processing.

REQUEST A DEMO TODAY

www.compilerworks.com