EASY ORCHESTRATION AND OBSERVABILITY IMPROVE ABILITY TO DELIVER VALUE

Cox Automotive's enterprise data services team maintains a data platform that primarily serves internal customers spanning across business units, though they also maintain a few data feeds to third parties. The enterprise data services team collects data from multiple internal sources and business units. "We use Databricks Workflows as our default orchestration tool to perform ETL and enable automation for about 300 jobs, of which approximately 120 are scheduled to run regularly," says Robert Hamlet, Lead Data Engineer, Enterprise Data Services, at Cox Automotive.

Jobs may be conducted weekly, daily or hourly. The amount of data processed in production pipelines today is approximately 720GB per day. Scheduled jobs pull from different areas both within and outside of the company. Hamlet uses Databricks Workflows to deliver data to the data science team, to the in-house data reporting team through Tableau, or directly into Power BI. "Databricks Workflows has a great user interface that allows you to quickly schedule any type of workflow, be it a notebook or JAR," says Hamlet. "Parametrization has been especially useful. It gives us clues as to how we can move jobs across environments. Workflows has all the features you would want from an orchestrator."

Hamlet also likes that Workflows provides observability into every workflow run and failure notifications so they can get ahead of issues quickly and troubleshoot before the data science team is impacted. "We use the job notifications feature to send failure notifications to a webhook, which is linked to our Microsoft Teams account," he says. "If we receive an alert, we go into Databricks to see what's going on. It's very useful to be able to peel into the run logs and see what errors occurred. And the Repair Run feature is nice to remove blemishes from your perfect history."

UNITY CATALOG AND DELTA SHARING IMPROVE DATA ACCESS ACROSS TEAMS

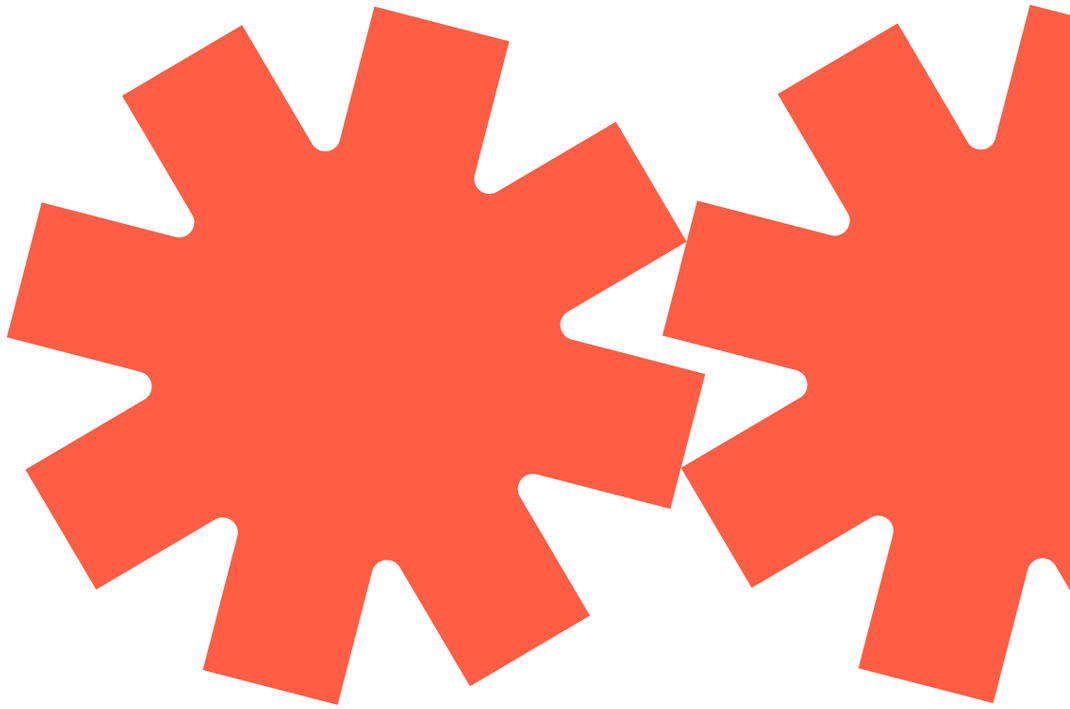
Hamlet's team recently began using Unity Catalog to manage data access, improving their existing method, which lacked granularity and was difficult to manage. "With our new workspace, we're trying to use more DevOps principles, infrastructure-as-code and groups wherever possible," he says. "I want to easily manage access to a wide range of data to multiple different groups and entities, and I want it to be as simple as possible for my team to do so. Unity Catalog is the answer to that."

The enterprise data services team also uses Delta Sharing, which natively integrates with Unity Catalog and allows Cox to centrally manage and audit shared data outside the enterprise data services team while ensuring robust security and governance. "Delta Sharing makes it easy to securely share data with business units and subsidiaries without copying or replicating it," says Hamlet. "It enables us to share data without the recipient having an identity in our workspace."

LOOKING AHEAD: INCORPORATING ADDITIONAL DATA INTELLIGENCE PLATFORM CAPABILITIES

Going forward, Hamlet plans to use Delta Live Tables (DLT) to make it easy to build and manage batch and streaming data pipelines that deliver data on the Databricks Data Intelligence Platform. DLT will help data engineering teams simplify ETL development and management. Eventually, Hamlet may also use Delta Sharing to easily share data securely with external suppliers and partners while meeting security and compliance needs. "DLT provides us an opportunity to make it simpler for our team. Scheduling Delta Live Tables will be another place we'll use Workflows," he says.

Hamlet is also looking forward to using the data lineage capabilities within Unity Catalog to provide his team with an end-to-end view of how data flows in the lakehouse for data compliance requirements and impact analysis of data changes. "That's a feature I'm excited about," Hamlet says. "Eventually, I hope we get to a point where we have all our data in the lakehouse, and we get to make better use of the tight integrations with things like data lineage and advanced permissions management."



Block — building a world-class data platform

Block standardizes on Delta Live Tables to expand secure economic access for millions

90%
Improvement in development velocity

150
Pipelines being onboarded in addition to the 10 running daily

INDUSTRY
Financial Services

PLATFORM
Delta Live Tables, Data Streaming, Machine Learning, ETL

CLOUD
AWS

Block is a global technology company that champions accessible financial services and prioritizes economic empowerment. Its subsidiaries, including Square, Cash App and TIDAL, are committed to expanding economic access. By utilizing artificial intelligence (AI) and machine learning (ML), Block proactively identifies and prevents fraud, ensuring secure customer transactions in real time. In addition, Block enhances user experiences by delivering personalized recommendations and using identity resolution to gain a comprehensive understanding of customer activities across its diverse services. Internally, Block optimizes operations through automation and predictive analytics, driving efficiency in financial service delivery. Block uses the Data Intelligence Platform to bolster its capabilities, consolidating and streamlining its data, AI and analytics workloads. This strategic move positions Block for the forthcoming automation-driven innovation shift and solidifies its position as a pioneer in AI-driven financial services.

ENABLING CHANGE DATA CAPTURE FOR STREAMING DATA EVENTS ON DELTA LAKE

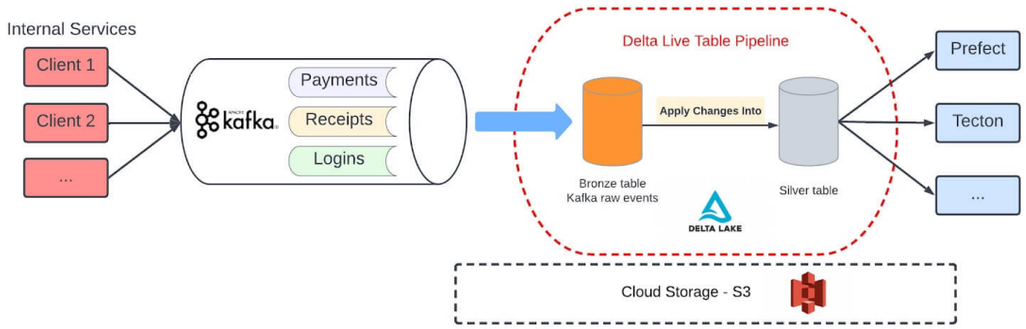
Block's Data Foundations team is dedicated to helping its internal customers aggregate, orchestrate and publish data at scale across the company's distributed system to support business use cases such as fraud detection, payment risk evaluation and real-time loan decisions. The team needs access to high-quality, low-latency data to enable fast, data-driven decisions and reactions.

Block had been consolidating on Kafka for data ingestion and Delta Lake for data storage. More recently, the company sought to make real-time streaming data available in Delta Lake as Silver (cleansed and conformed) data for analytics and machine learning. It also wanted to support event updates and simple data transformations and enable data quality checks to ensure higher-quality data. To accomplish this, Block considered a few alternatives, including the Confluent-managed Databricks Delta Lake Sink connector, a fully managed solution with low latency. However, that solution did not offer change data capture support and had limited transformation and data quality check support. The team also considered building their own solution with Spark Structured Streaming, which also provided low latency and strong data transformation capabilities. But that solution required the team to maintain significant code to define task workflows, change data and capture logic. They'd also have to implement their own data quality checks and maintenance jobs.

LEVERAGING THE LAKEHOUSE TO SYNC KAFKA STREAMS TO DELTA TABLES IN REAL TIME

Rather than redeveloping its data pipelines and applications on new, complex, proprietary and disjointed technology stacks, Block turned to the Data Intelligence Platform and Delta Live Tables (DLT) for change data capture and to enable the development of end-to-end, scalable streaming pipelines and applications. DLT pipelines simply orchestrate the way data flows between Delta tables for ETL jobs, requiring only a few lines of declarative code. It automates much of the operational complexity associated with running ETL pipelines and, as such, comes with preselected smart defaults yet is also tunable, enabling the team to optimize and debug easily. "DLT offers declarative syntax to define a pipeline, and we believed it could greatly improve our development velocity," says Yue Zhang, Staff Software Engineer for the Data Foundations team at Block. "It's also a managed solution, so it manages the maintenance tasks for us, it has data quality support, and it has advanced, efficient autoscaling and Unity Catalog integration."

Today, Block's Data Foundations team ingests events from internal services in Kafka topics. A DLT pipeline consumes those events into a Bronze (raw data) table in real time, and they use the DLT API to apply changes and merge data into a higher-quality Silver table. The Silver table can then be used by other DLT pipelines for model training, to schedule model orchestration, or to define features for a features store. "It's very straightforward to implement and build DLT pipelines," says Zhang.



The Block Data Foundations team's streaming data architecture with Delta Live Tables pipelines

Using the Python API to define a pipeline requires three steps: a row table, a Silver table and a merge process. The first step is to define the row table. Block consumes events from Kafka, performs some simple transformations, and establishes the data quality check and its rule. The goal is to ensure all events have a valid event ID.

The next step is to define the Silver table or target table, its storage location and how it is partitioned. With those tables defined, the team then determines the merge logic. Using the DLT API, they simply select APPLY CHANGES INTO. If two units have the same event ID, DLT will choose the one with the latest ingest timestamp. "That's all the code you need to write," says Zhang.

Finally, the team defines basic configuration settings from the DLT UI, such as characterizing clusters and whether the pipeline will run in continuous or triggered modes.

Following their initial DLT proof of concept, Zhang and his team implemented CI/CD to make DLT pipelines more accessible to internal Block teams. Different teams can now manage pipeline implementations and settings in their own repos, and, once they merge, simply use the Databricks pipelines API to create, update and delete those pipelines in the CI/CD process.

BOOSTING DEVELOPMENT VELOCITY WITH DLT

Implementing DLT has been a game-changer for Block, enabling it to boost development velocity. "With the adoption of Delta Live Tables, the time required to define and develop a streaming pipeline has gone from days to hours," says Zhang.

Meanwhile, managed maintenance tasks have resulted in better query performance, improved data quality has boosted customer trust, and more efficient autoscaling has improved cost efficiency. Access to fresh data means Block data analysts get more timely signals for analytics and decision-making, while Unity Catalog integration means they can better streamline and automate data governance processes. "Before we had support for Unity Catalog, we had to use a separate process and pipeline to stream data into S3 storage and a different process to create a data table out of it," says Zhang. "With Unity Catalog integration, we can streamline, create and manage tables from the DLT pipeline directly."

Block is currently running approximately 10 DLT pipelines daily, with about two terabytes of data flowing through them, and has another 150 pipelines to onboard. "Going forward, we're excited to see the bigger impacts DLT can offer us," adds Zhang.

Trek — global bicycle leader accelerates retail analytics

80%-90%
Acceleration in runtime of retail analytics solution globally

3X
Increase in daily data refreshes on Databricks

1 Week
Reduction in ERP data replication, which now happens in near real time

"How do you scale up analytics without blowing a hole in your technology budget? For us, the clear answer was to run all our workloads on Databricks Data Intelligence Platform and replicate our data in near real-time with Qlik."

— Garrett Baltzer, Software Architect, Data Engineering, Trek Bicycle

INDUSTRY
Retail and Consumer Goods

SOLUTION
Real-Time Point-of-Sale Analytics

PLATFORM
Delta Lake, Databricks SQL, Delta Live Tables, Data Streaming

PARTNER
Qlik

CLOUD
Azure

Trek Bicycle started in a small Wisconsin barn in 1976, but their founders always saw something bigger. Decades later, the company is on a mission to make the world a better place to live and ride. Trek only builds products they love and provides incredible hospitality to customers as they aim to change the world for the better by getting more people on bikes. Frustrated by the rising costs and slow performance of their data warehouse, Trek migrated to Databricks Data Intelligence Platform. The company now uses Qlik to replicate their ERP data to Databricks in near real-time and stores data in Delta Lake tables. With Databricks and Qlik, Trek has dramatically accelerated their retail analytics to provide a better experience for their customers with a unified view of the global business to their data consumers, including business and IT users.

databricks

SLOW DATA PROCESSING HINDERS RETAIL ANALYTICS

As Trek Bicycle works to make the world better by encouraging more people to ride bikes, the company keeps a close eye on what's happening in their hundreds of retail stores. But until recently, running analytics on their retail data proved challenging because Trek relied on a data warehouse that couldn't scale cost-effectively.

"The more stores we added, the more information we added to our processes and solutions," explained Garrett Baltzer, Software Architect, Data Engineering, at Trek Bicycle. "Although our data warehouse did scale to support greater data volumes, our processing costs were skyrocketing, and processes were taking far too long. Some of our solutions were taking over 30 hours to produce analytics, which is unacceptable from a business perspective."

Adding to the challenge, Trek's data infrastructure hindered the company's efforts to achieve a global view of their business performance. Slow processing speeds meant Trek could only process data once per day for one region at a time.

"We were processing retail data separately for our North American, European and Asia-Pacific stores, which meant everyone downstream had to wait for actionable insights for different use cases," recalled Advait Raje, Team Lead, Data Engineering, at Trek Bicycle. "We soon made it a priority to migrate to a unified data platform that would produce analytics more quickly and at a lower cost."

DELTA LAKE UNIFIES RETAIL DATA FROM AROUND THE GLOBE

Seeking to modernize their data infrastructure to speed up data processing and unify all their data from global sources, Trek started migrating to the Databricks Data Intelligence Platform in 2019. The company's processing speeds immediately increased. Qlik's integration with the Databricks Data Intelligence Platform helps feed Trek's lakehouse. This replication allows Trek to build a wide range of valuable data products for their sales and customer service teams.

"Qlik enabled us to move relevant ERP data into Databricks where we don't have to worry about scaling vertically because it automatically scales parallel. Since 70 to 80% of our operational data comes from our ERP system, Qlik has made it possible to get far more out of our ERP data without increasing our costs," Baltzer explained.

Trek is now running all their retail analytics workloads in the Databricks Data Intelligence Platform. Today, Trek uses the Databricks Data Intelligence Platform to collect point-of-sale data from nearly 450 stores around the globe. All computation happens on top of Trek's lakehouse. The company runs a semantic layer on top of this lakehouse to power everything from strategic high-level reporting for C-level executives to daily sales and operations reports for individual store employees.

"Databricks Data Intelligence Platform has been a game-changer for Trek," said Raje. "With Qlik Cloud Data Integration on Databricks, it became possible to replicate relevant ERP data to our Databricks in real time, which made it far more accessible for downstream retail analytics. Suddenly, all our data from multiple repositories was available in one place, enabling us to reduce costs and deliver on business needs much more quickly."

Trek's BI and data analysts leverage Databricks SQL, their serverless data warehouse, for ad hoc analysis to answer business questions much more quickly. Internal customers can leverage Power BI connecting directly to Databricks to consume retail analytics data from Gold tables. This ease of analysis helps the company monitor and enhance their Net Promoter Scores. Trek uses Structured Streaming and Auto Loader functionality within Delta Live Tables to transform the data from Bronze to Silver or Gold, according to the medallion architecture.

"Delta Live Tables have greatly accelerated our development velocity," Raje reported. "In the past, we had to use complicated ETL processes to take data from raw to parsed. Today, we just have one simple notebook that does it, and then we use Delta Live Tables to transform the data to Silver or Gold as needed."

DATA INTELLIGENCE PLATFORM ACCELERATES ANALYTICS BY 80% TO 90%

By moving their data processing to the Databricks Data Intelligence Platform and integrating data with Qlik, Trek has dramatically increased the speed of their processing and overall availability of data. Prior to implementing Qlik, they had a custom program that, once a week, on a Sunday, replicated Trek's ERP data from on-premises servers to a data lake using bulk copies. Using Qlik, Trek now replicates relevant data from their ERP system as Delta tables directly in their lakehouse.

"We used to work with stale ERP data all week because replication only happened on Sundays," Raje remarked. "Now we have a nearly up-to-the-minute view of what's going on in our business. That's because Qlik lets us keep replicating through the day, streaming data from ERP into our lakehouse."

Trek's retail analytics solution used to take 48+ hours to produce meaningful results. Today, Trek runs the solution on the Databricks Data Intelligence Platform to get results in six to eight hours — an 80 to 90% improvement, thus allowing daily runs. A complementary retail analytics solution went from 12–14 hours down to under 4–5 hours, thereby enabling the lakehouse to refresh three times per day, compared to only once a day previously.

"Before Databricks, we had to run our retail analytics once a day on North American time, which meant our other regions got their data late," said Raje. "Now, we refresh the lakehouse three times per day, one for each region, and stakeholders receive fresh data in time to drive their decisions. Based on the results we've achieved in the lakehouse, we're taking a Databricks-first approach to all our new projects. We're even migrating many of our on-premises BI solutions to Databricks because we're all-in on the lakehouse."

"Databricks Data Intelligence Platform, along with data replication to Databricks using Qlik, aligns perfectly with our broader cloud-first strategy," said Steve Novoselac, Vice President, IT and Digital, at Trek Bicycle. "This demonstrates confidence in the adoption of this platform at Trek."

Coastal Community Bank — mastering the modern data platform for exponential growth

< 10
Minutes to securely share large datasets across organizations

99%
Decrease in processing duration (2+ days to 30 min.)

12X
Faster partner onboarding by eliminating sharing complexity

"We've done two years' worth of work here in nine months. Databricks enables access to a single source of truth and our ability to process high volumes of transactions that gives us confidence we can drive our growth as a community bank and a leading banking-as-a-service provider."

— Curt Queyrouze, President, Coastal Community Bank

Many banks continue to rely on decades-old, mainframe-based platforms to support their back-end operations. But banks that are modernizing their IT infrastructures and integrating the cloud to share data securely and seamlessly are finding they can form an increasingly interconnected financial services landscape. This has created opportunities for community banks, fintechs and brands to collaborate and offer customers more comprehensive and personalized services. Coastal Community Bank is headquartered in Everett, Washington, far from some of the world's largest financial centers. The bank's CCBX division offers banking as a service (BaaS) to financial technology companies and broker-dealers. To provide personalized financial products, better risk oversight, reporting and compliance, Coastal turned to the Databricks Data Intelligence Platform and Delta Sharing, an open protocol for secure data sharing, to enable them to share data with their partners while ensuring compliance in a highly regulated industry.

INDUSTRY
Financial Services

SOLUTION
Financial Crimes Compliance, Customer Profile Scoring, Financial Reconciliation, Credit Risk Reporting, Synthesizing Multiple Data Sources, Data Sharing and Collaboration

PLATFORM
Delta Lake, ETL, Delta Sharing, Data Streaming, Databricks SQL

CLOUD
Azure

databricks