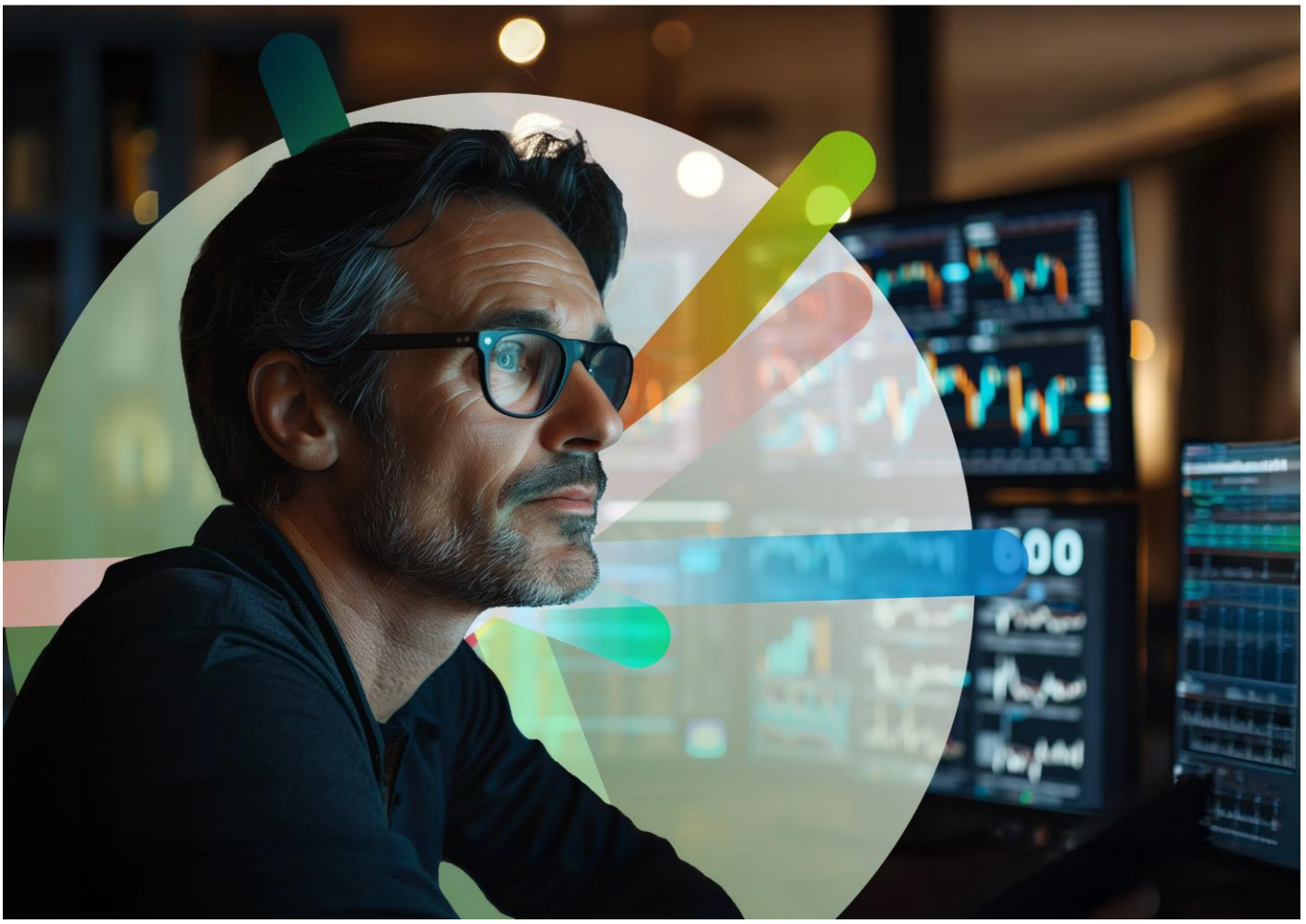
Deloitte. | databricks



Unlocking the Power of Databricks Lakeflow and Transforming Data Engineering

Explore Databricks Lakeflow's transformative power and its critical role within the Databricks Data Intelligence Platform, its integration into a Data and Analytics Platform (DAP), and the extensive benefits it offers to the Databricks ecosystem.

Traditional Extraction Challenges

Let's take a moment to reflect on the evolution of data integration. In the past, database connections (using the timestamp) were the stalwarts of data transfer, reliably facilitating data movement between systems. However, these once-trusted methods faced challenges as data volumes and complexity increased. Moreover, the growing demand for real-time data insights has further accelerated the need for more robust, scalable, and agile solutions that can handle the speed and volume of today's data landscape.

Enter Change Data Capture (CDC) technologies—the new superheroes of the data world! They swoop in to save the day with real-time data processing. Before diving into Lakeflow, understanding the CDC's value is key. It's like having a superpower for your data!

Traditional Extraction Challenges

Traditional data extraction is like filling a swimming pool with a coffee mug—slow and inefficient.

No Timestamp? More Complexity

When tables lack timestamps, the entire dataset must be extracted into a staging area to identify changes. This process requires more computing resources, and the computing needs increase with the volume of sources.

Overhead on Source Systems

Extraction logic creates significant overhead*, diverting resources from critical applications. This slows source systems, hindering operational demands. Therefore, extractions are often limited to off-peak hours.

Delayed Data Availability

Even if extractions occur during off-peak hours, the data may not reach business users quickly enough, as it can be several hours old when available.

Operational Analytics Limitations

Some businesses set up replicated databases on-premises to run operational queries without impacting the source system. However, this approach is limited to operational analytics, doesn't easily integrate with cloud-based data, and becomes a scalability bottleneck as analytics needs grow.

Limited Visibility

Traditional extraction methods only provide the latest data, not the series of actions leading up to it. For instance, tracking customer purchase behavior before a purchase is crucial for retailers to understand which items were added and removed from carts.

Multiplying the Load

Every user needing data from the source must tap into it, multiplying the issue and load on the source system.

Failed Extracts

If an extraction fails, it requires another trip to the source system, re-extracting the same data, which the source system may not appreciate.

Multiple data Sources

There are data sources that are spread across multiple clouds and on-premise systems, which result in specialized team dependencies, lack of governance, and inefficient development.

** Not all sources support native Change Data Capture (CDC), so extraction will often need to rely on cursors, especially for many sources like SaaS applications.*

Change Data Capture (CDC) Technology

A Better Approach: Capture & Replicate Database Changes in Real-Time

What if there was a way to extract data with minimal impact on the source system while providing real-time data access, highly performant and scalable analytical data storage, and ensuring that consumers don't need to tap into the source database repeatedly?

Databases store all actions—updates, inserts, and deletes—in transactional logs used by database management systems but separate from the database itself. By tapping into these logs, we can read the actions with minimal impact, akin to a "pinch" rather than a "punch." This approach eliminates the issues mentioned above.

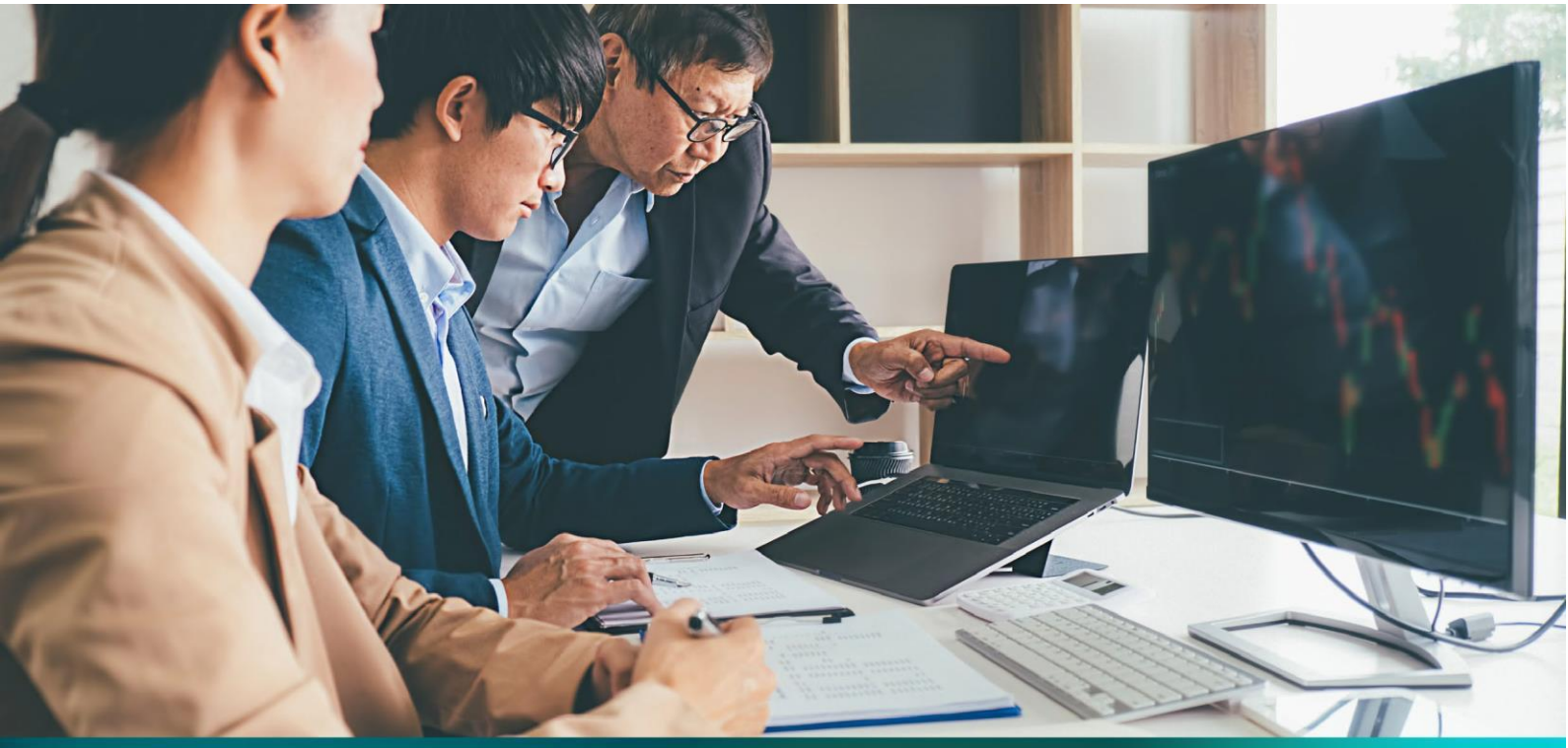
Databricks Lakeflow

Lakeflow: a unified, intelligent solution for data engineering. Built on the Data Intelligence Platform, Lakeflow covers ingestion, transformation, and orchestration.

Lakeflow is designed to provide a comprehensive platform for ingesting data from various sources. It supports a wide range of users, including data engineers, data scientists, data analysts, and other professionals, enabling them to orchestrate workloads involving data, AI, and SQL. Key features include unified governance with Unity Catalog, a self-service interface for all practitioners, and an efficient, managed solution. Databricks Lakeflow addresses modern data challenges, enhancing productivity across your organization.

Lakeflow Connect

Databricks Lakeflow Connect leverages log-based optimal Change Data Capture (CDC), which accesses the source database's transaction logs, and other techniques to capture data changes and DML operations like inserts, updates, and deletes. Additionally, when CDC is not available on a source, a query-based method leveraging a cursor column for incremental ingestion will soon be available. * Note – Lakeflow Connect is well-suited for capturing change data. For automatically handling change data in your pipelines, Lakeflow Declarative Pipelines offers built-in support for change data capture with AutoCDC: https://docs.databricks.com/aws/en/dlt/cdc



Lakeflow provides benefits for ingestion and transformation:

Minimal Performance Impact

Lakeflow Connect introduces minimal overhead to the source system, ensuring business-critical applications remain unaffected.

Internal Data Storage

Once data is read from the logs, it's stored internally, so subsequent requests don't need to revisit the source system, further reducing the load.

Checkpointing for Replication Failures

If replication fails, Lakeflow keeps track of each commit (checkpoint), allowing it to resume from the point of failure upon restart.

Target Database Synchronization

Lakeflow ensures the target database remains in sync with the source system, applying updates, inserts, and deletes as needed.

Detailed Action Visibility

Users can access not just the synchronized data but also the detailed actions that occurred in the source system.

Multi-Target Delivery

Lakeflow Declarative Pipelines can sink data to multiple targets, including Kafka, external Delta tables, and more

Note: Some configurations and setups may vary based on the specific data sources.

What sets Lakeflow Connect apart from third-party CDC tools?

I want to share a personal anecdote that effectively illustrates the point. Drawing from our own experience, we can provide a relatable example highlighting the key issue.

Consider a scenario where a critical system component fails, prompting the consideration of a third-party alternative. Initially, this option may seem appealing due to its cost-effectiveness and additional features. However, upon closer examination, it becomes apparent that such solutions often come with integration challenges. While they may excel in certain areas, they frequently require workarounds to function seamlessly within a specific ecosystem. This trade-off is common when adopting solutions not initially designed for a particular environment, highlighting the importance of carefully evaluating compatibility and integration requirements.

Third-party CDC tools have played a crucial role in filling critical data integration gaps, offering valuable solutions for many organizations. They've paved the way for efficient data replication and are instrumental in many data architectures. With the introduction of Databricks Lakeflow Connect, we have a set of built-in, managed connectors that seamlessly integrate with the Databricks Data Intelligence Platform. It's like choosing between a well- respected aftermarket car stereo and a manufacturer's premium audio system - both have their merits, and now you have a better choice based on your specific needs and existing infrastructure.

Traditionally, Change Data Capture (CDC) tools have focused on connecting to databases to extract real-time data. However, the need for real-time data now extends far beyond databases, encompassing sources like Google Ads, Google Analytics, Salesforce, and ServiceNow. To address this challenge, Lakeflow Connect Connectors provides a streamlined solution, offering the ability to integrate these diverse data sources without the complexity of custom-built connectors. This simplifies the process, making it easier to access and leverage critical data from various applications.

Extensive Source Support: Lakeflow Connect connects to a wide range of data sources, including **SQL Server, Salesforce, Workday, Google Analytics, ServiceNow, with expanding support for MySQL, Postgres, Oracle, NetSuite, and Dynamics 365.**

But here's the twist: Lakeflow offers additional perks beyond CDC capabilities. Let's explore these extras to help you decide whether an integrated product or a third-party tool is right for you.

So far, we've explored Lakeflow Connect and its powerful CDC capabilities that forms the solution's foundation. Let's examine how Lakeflow extends beyond simple data replication to deliver transformative value through its additional functionalities.

Benefits Beyond CDC Replication and Pushing the Limits

Pipelines

Advanced Transformations

Traditional CDC tools are limited when it comes to complex data transformations. They typically allow only minor adjustments and lack the capability to handle more intricate processes. In contrast, Lakeflow enables the creation of sophisticated pipelines using SQL and Python. It also supports real-time mode for Apache Spark, allowing stream processing with significantly faster latency than micro-batch processing.

Serverless and Governance

Lakeflow leverages serverless computing and unified governance through Databricks Unity Catalog. This is particularly important because it provides complete Unity Catalog governance capabilities and an end-to-end view without incurring additional costs for Unity Catalog. Furthermore, serverless computing enhances Lakeflow's efficiency by offering cost-effective scalability, reduced operational overhead, and improved reliability, allowing organizations to focus on data-driven insights rather than infrastructure management.

Accomplishing Complex Streaming and Batch Processing

Built on the core declarative pipeline framework (which has now been open-sourced as Spark Declarative Pipelines), Lakeflow Declarative Pipelines enables you to develop sophisticated transformations using SQL and Python. For example, you can implement complex data quality rules, perform multi-table joins with historical context, or apply machine learning transformations—all within the same pipeline. Databricks automates data orchestration, incremental processing, and compute infrastructure autoscaling, so you don't need to worry about these tasks. Additionally, it monitors data quality and allows you to choose between streaming and batch processing without requiring separate codebases. Please refer to our earlier blog post for more details on Lakeflow Declarative Pipelines and its advantages.

While many third-party CDC tools excel at their core data replication functionality, integrating them into a comprehensive transformation framework requires additional engineering effort. Lakeflow's built-in integration with Lakeflow Declarative Pipelines represents a complementary approach that builds upon the foundation established by these proven technologies. Now, what's the third benefit.

Orchestration and Integration Within an Enterprise Ecosystem

CDC transformations are not standalone processes but part of a broader enterprise data platform ecosystem. Your applications depend on ingestion jobs, notebook executions, SQL queries, AI/ML training, deployments, and more. Can all these jobs be orchestrated with appropriate dependencies? Databricks offers an out-of-the-box solution called Databricks Lakeflow Jobs, which addresses this need. This is the third benefit that rounds out Lakeflow's capabilities.

Collective Comprehensive Solutions

In summary, Lakeflow Connect, Lakeflow Declarative Pipelines, and Lakeflow Jobs provide enterprises with the necessary capabilities under one roof.

This includes native connections for the highest performance, best value for investment, lower total cost of ownership, complete governance, observability, monitoring, data security, and access through Databricks Data Intelligence Platform. Like all new products, Lakeflow enhancements move through various release states to general availability, ensuring a reliable outcome.

By integrating Lakeflow, organizations maximize their current data infrastructure investments and acquire new capabilities to handle evolving data demands. By leveraging Databricks' integrated solution, they can innovate faster, develop unprecedented data solutions, and drive meaningful business outcomes.

The Deloitte and Databricks Advantage

To meet today's advanced analytics demands, you need a faster path to making smarter, data-driven decisions. By providing the necessary tools, people, and accelerators, coupled with vast experience in implementing and scaling the most innovative technologies, Deloitte and Databricks deliver elevated capabilities, accelerated efficiency, and proven experience. Together, Deloitte and Databricks can transform your data systems from siloed and complex to unified, efficient, and cost-effective. That means more time better spent leveraging the insights that help you successfully run your business.

Ready to transform your data and analytics platform with Lakeflow

Mani Kandasamy
Databricks Alliance CTO, Deloitte Consulting LLP
mkandasamy@deloitte.com

Vishal Vibhandik
Partner Solution Architect, Databricks
vishal.vibhandik@databricks.com