Data Migration to Databricks

Automated migration and transformation to Delta Lake on Databricks for fastest time to AI and ML

Cloud-native analytic and machine learning environments like Databricks hold large potential value for organizations that have been constrained by their on-premises infrastructure. It is not just a case of migrating them in a "lift and shift" manner, but of identifying and taking advantage of the opportunities for modernizing those environments at the same time.

Cirata's unique Data approach to migrating data at scale without disrupting the use of those datasets while an organization adopts cloud infrastructure and services has been a critically-important answer to these challenges.

From migration to modernization

The combination of Cirata and Databricks enables organizations to:

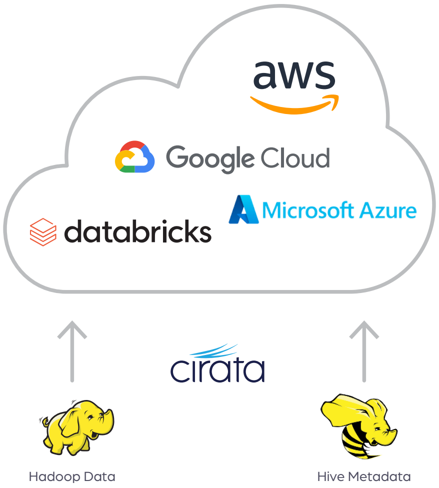
1 Automate their Hadoop data and Hive metadata migration with zero downtime and zero business disruption

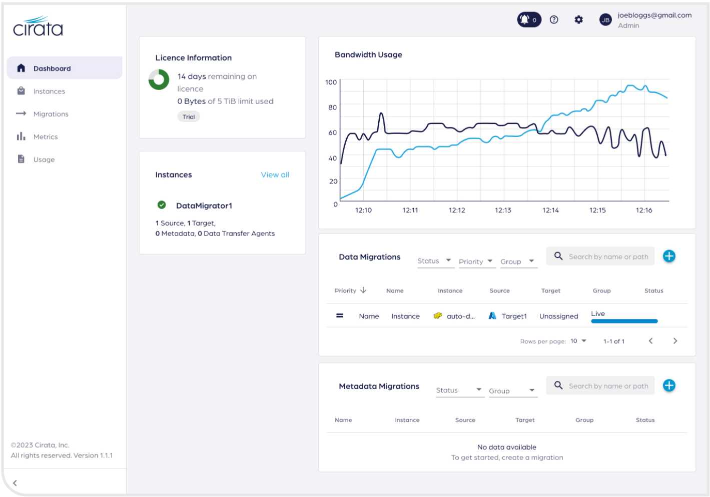
2 Modernize their data architecture with a unified analytics platform that ensures data reliability and data consistency

Automate data and metadata migration to Databricks

Cirata Data Migrator is a safe and reliable cloud migration solution that automates the migration of Hadoop data and Hive metadata to the cloud.

Data Migrator deployment is performed in minutes, and requires no changes to applications or normal business operations. Migrations of any scale can begin immediately and be performed while the source data is under active change without requiring any production system downtime or business disruption, and with zero risk of data loss.





Licence Information

14 days remaining on licence

0 Bytes of 5 TiB limit used

Trial

Instances View all

DataMigrator1

1 Source, 1 Target,
0 Metadata, 0 Data Transfer Agents

Data Migrations Status Priority Group Search by name or path

Priority	Name	Instance	Source	Target	Group	Status
=	Name	Instance	auto-d...	Target1	Unassigned	Live

Rows per page: 10 1-1 of 1

Metadata Migrations Status Group Search by name or path

Name	Instance	Source	Target	Group	Status

No data available
To get started, create a migration

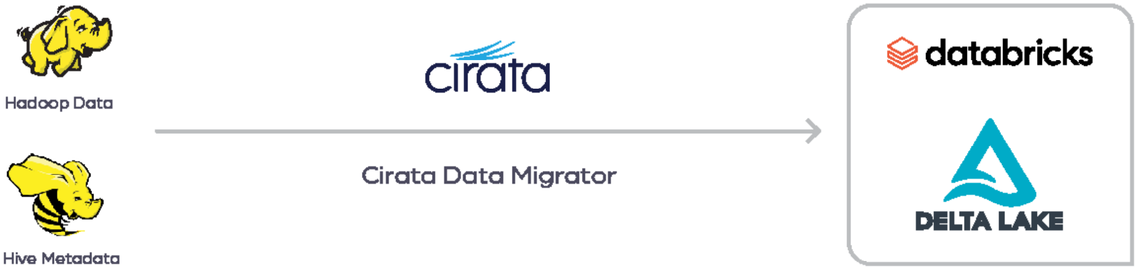
©2023 Cirata, Inc.
All rights reserved. Version 1.1.1

Data Migrator provides three key Databricks–specific functionalities:

- Make Apache Hive metadata available directly in Databricks workspaces using live migration so that ongoing changes to source metadata are reflected immediately in the Databricks target.
- Transform the on–premises data formats used in Hadoop and Hive to the Databricks–preferred Delta Lake form, so that users can take full advantage of the features that are unique to the combination of Databricks and Delta Lake.
- Support for unity catalog to ensure organizations can define the catalog as part of the migrations so data can be made readily available for end users in correct catalog locations for immediate processing.

By combining data and metadata, and by making on–premises content immediately usable in an ideal form in Databricks, migration tasks that previously required constructing data pipelines to transform, filter and adjust data, as well as significant up–front planning for staging and processing work are now eliminated. Equally, work that would otherwise be required for setting up auto–load pipelines to attempt to identify newly–landed data, and convert it to a final form as part of a processing pipeline can be set aside.

Data Migrator is in control of when datasets land in the cloud so it can initiate the work required to load them into a final form, bypassing the need to detect newly–created data, or identify changes to existing data. It is more efficient, more scalable, and entirely automated.

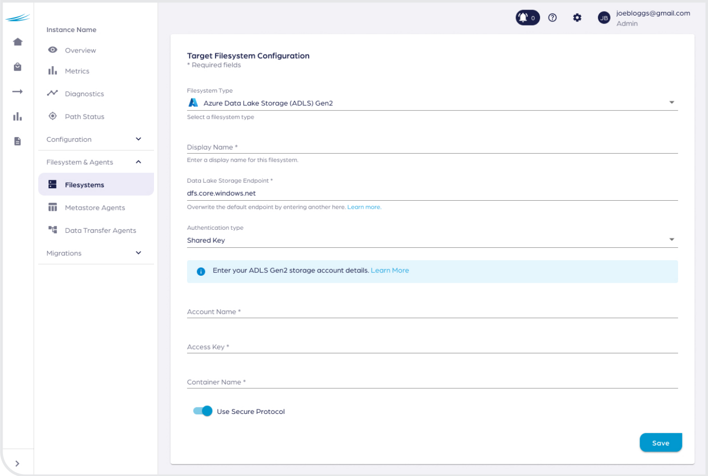


Hadoop data and Hive metadata migration to Databricks

Making Hive data and metadata available for direct use as Delta Lake content in Databricks with Data Migrator is a simple as four steps:

Step 1 – define your targets

Configure Data Migrator to have a data migration target available for your chosen cloud storage and for Databricks. Choose to convert content to the Delta Lake format when you create your Databricks metadata target.



Target Filesystem Configuration
* Required fields

Filesystem Type: Azure Data Lake Storage (ADLS) Gen2
Select a filesystem type

Display Name *
Enter a display name for this filesystem.

Data Lake Storage Endpoint *: dfs.core.windows.net
Overwrite the default endpoint by entering another here. Learn more.

Authentication type: Shared Key

Enter your ADLS Gen2 storage account details. Learn More

Account Name *

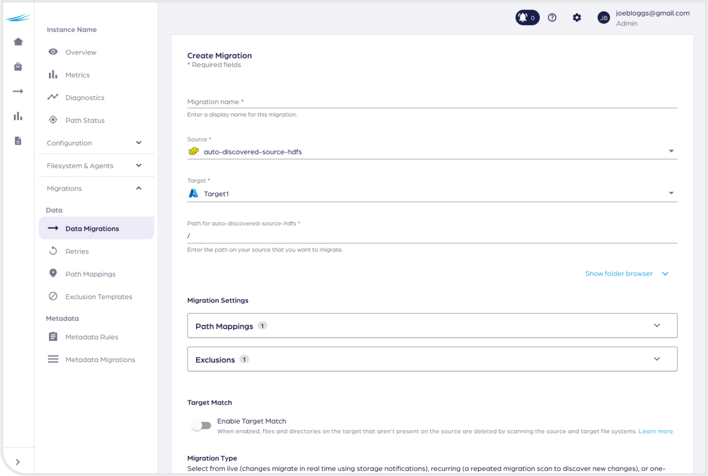
Access Key *

Container Name *

Use Secure Protocol

Step 2 – define your data migration rule

Select the data you want to migrate, and optionally any data that you may want to exclude from the migration. Select "auto start migration" to automatically begin migrating data.



Create Migration
* Required fields

Migration name *
Enter a display name for this migration.

Source *: auto-discovered-source-hdfs

Target *: Target1

Path for auto-discovered-source-hdfs *: /
Enter the path on your source that you want to migrate.

Migration Settings

Path Mappings 1

Exclusions 1

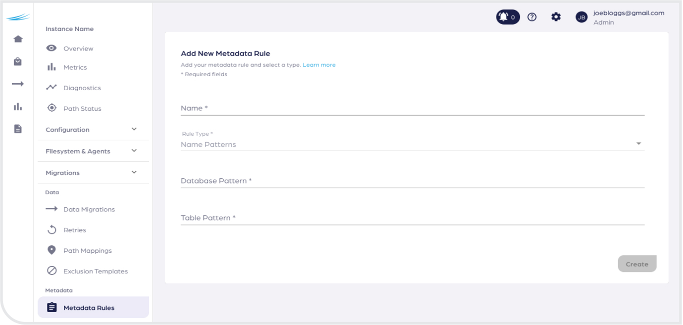
Target Match

Enable Target Match
When enabled, files and directories on the target that aren't present on the source are deleted by scanning the source and target file systems. Learn more

Migration Type
Select from live (changes migrate in real time using storage notifications), recurring (a repeated migration scan to discover new changes), or one–

Step 3 – define your metadata rules

Create and define the various rules to associate to your metadata migration. Name your rule and provide the database and table pattern associated to the metadata you want to migrate.



Add New Metadata Rule

Add your metadata rule and select a type. Learn more

* Required fields

Name *

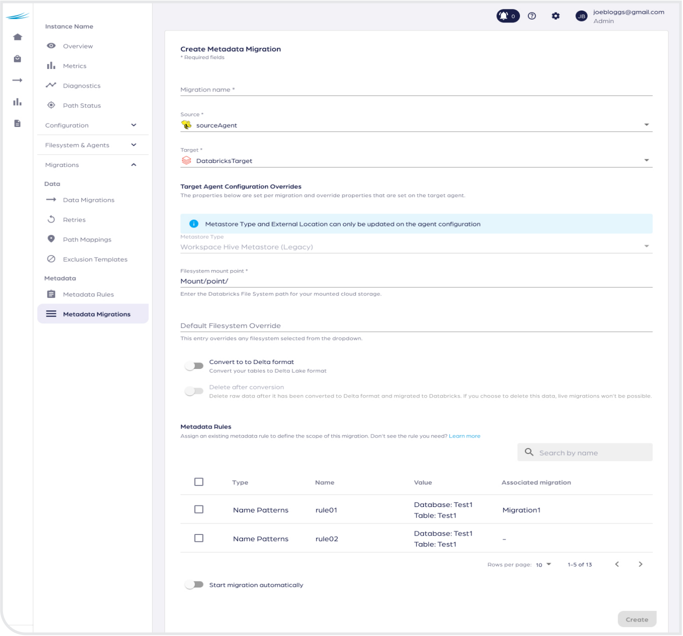
Rule Type *
Name Patterns

Database Pattern *

Table Pattern *

Step 4 – define your metadata migration

Create your metadata migration and choose to convert the metadata to delta format. Select the metadata rule(s) to apply to the migration and then "Start migration automatically" to automatically begin migrating metadata.



Create Metadata Migration

* Required fields

Migration name *

Source *
sourceAgent

Target *
DatabricksTarget

Target Agent Configuration Overrides

The properties below are set per migration and override properties that are set on the target agent.

Metastore Type and External Location can only be updated on the agent configuration

Metastore Type
Workspace Hive Metastore (Legacy)

Filesystem mount point *
Mount/point/

Enter the Databricks File System path for your mounted cloud storage.

Default Filesystem Override

This entry overrides any filesystem selected from the dropdown.

Convert to to Delta format
Convert your tables to Delta Lake format

Delete after conversion
Delete raw data after it has been converted to Delta format and migrated to Databricks. If you choose to delete this data, live migrations won't be possible.

Metadata Rules

Assign an existing metadata rule to define the scope of this migration. Don't see the rule you need? Learn more

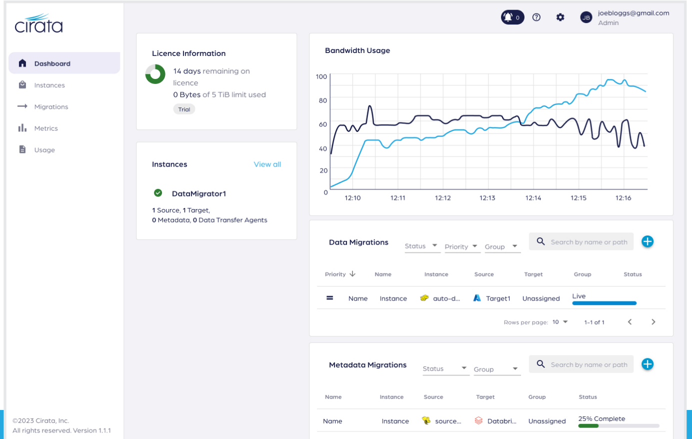
□	Type	Name	Value	Associated migration
□	Name Patterns	rule01	Database: Test1 Table: Test1	Migration1
□	Name Patterns	rule02	Database: Test1 Table: Test1	–

Rows per page: 10 1–5 of 13

Start migration automatically

That's it. Your data and metadata will immediately begin to migrate to Databricks.

You can monitor the migration progress in the Cirata UI and optionally select to receive alerts and notifications sent to you directly.



Licence Information

14 days remaining on licence
0 Bytes of 5 TiB limit used
Trial

Instances

DataMigrator1
1 Source, 1 Target,
0 Metadata, 0 Data Transfer Agents

Data Migrations

Priority	Name	Instance	Source	Target	Group	Status
=	Name	Instance	auto-d...	Target1	Unassigned	Live

Rows per page: 10 1–1 of 1

Metadata Migrations

Name	Instance	Source	Target	Group	Status
Name	Instance	source...	Databri...	Unassigned	25% Complete

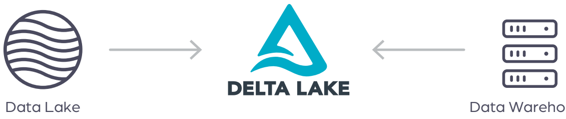
©2023 Cirata, Inc.
All rights reserved. Version 1.1.1

Modernize data architecture with a unified analytics platform

While the cloud brings efficiencies for data lakes there remains concerns about the reliability and the consistency of the data. Data Lakes typically have multiple data pipelines reading and writing data concurrently, and data engineers have to go through a tedious process to ensure data integrity, due to the lack of transactions.

Databricks handles with ease all of the analytic processes that previously suffered under inflexible and cumbersome Hadoop deployments on-premises.

Its success is prompting enterprises like yours to consider migrating on-premises Hadoop workloads to the cloud, and Databricks particularly. Databricks provides an elegant answer to the concern around the sunk costs of skills, infrastructure and systems built on-premises in Hadoop by allowing the same technologies, applications and systems to operate without change either on-premises or in the cloud. Delta Lake provides the storage layer on top of your existing storage to support enterprise workloads across streaming and batch requirements to better manage data lakes at scale. Delta Lake supports multiple simultaneous readers and writers for mixed batch and streaming data making it easy for data teams to run interactive queries and batch historic backfill out of the box.



Unity Catalog provides centralized access control, auditing, lineage, and data discovery capabilities across Databricks workspaces.

Databricks provides a Unified Analytics Platform powered by Apache Spark for data science teams to collaborate with data engineering and lines of business to build data products. You can achieve faster time-to-value with Databricks by creating analytic workflows that go from ETL and interactive exploration to production. Databricks also makes it easier for you to focus on your data rather than hardware by providing a fully managed, scalable, and secure cloud infrastructure that reduces operational complexity and total cost of ownership.

The Cirata and Databricks outcome

For the first time, organizations that want to migrate on-premises Hadoop content from Hive to Databricks can do so at scale, for live data, automatically and selectively without any disruption or change to their existing systems. Data Migrator includes capabilities to target Databricks Unity Catalog and to transform from HDFS and Hive-specific data formats to the Delta Lake form.

This means that cloud migration and modernization are within the scope of constrained teams, without risk, and without imposing a big-bang cutover. Workloads and data that have been locked-up on-premises can now be used immediately in the cloud, using the modern data analytics platform offered by Databricks.

Data Migrator capabilities

- **Quick deployment and operation:** Data Migrator is installed on an edge node of your Hadoop cluster. Deployment can be performed in minutes without impact to current operations, so users can begin migrations immediately.
- **Self-service user experience:** Migrations are designed to be easy to configure and perform, requiring simple definition of your target environment and full control of exactly what data to migrate and what data to exclude.
- **Complete and continuous migration:** Migrates existing data sets with a single pass through the source storage system, eliminating the overhead of repeated scans, while also supporting continuous migration of any ongoing changes from source to target with zero disruption to current production systems.
- **Hadoop data and Hive metadata migration:** Supports migration of HDFS data and Hive metadata to public cloud, as well as to other on-premises environments.
- **Multiple source and target systems support:** Supports HDFS distributions v2.6 and higher as source systems, and all leading cloud service providers and other select ISVs such as Databricks as the target systems.
- **Migration at any scale:** Migrates big data sets at any scale, from terabytes to multi petabytes, without impact to current production environments. Begin risk free for small migrations and scale up to multi petabyte initiatives without needing any additional installation requirements.
- **Browser-based user interface:** Users can leverage the Cirata UI, a browser-based user interface that allows them to manage the full data migration (data and metadata) from the single management console.
- **Programmatic interface:** Migrations can also be managed through a comprehensive and intuitive command-line interface or using the self-documenting REST API to integrate the solution with other programs as needed.
- **Configurability and control:** Ability to configure the migrations to meet the organizations specific needs. Including standard configuration such as defining sources, targets, and data to be migrated, as well as advanced capabilities such as path mapping and network bandwidth management controls.
- **Metrics and monitoring:** Information to keep you updated on the migration jobs, from health and status metrics providing estimates for migration completion to email notifications and real-time insights regarding usage and promote hands-off operations.

Delta Lake key features

ACID Transactions: Delta Lake brings ACID transactions to your data lakes. It provides serializability, the strongest level of isolation level, allowing you to build reliability into your data processing and analytics effortlessly.

Scalable Metadata: Delta Lake treats metadata just like data, leveraging Spark's distributed processing power to handle all its metadata. As a result, Delta Lake can handle petabyte-scale tables with billions of partitions and files at ease.

Data Versioning: Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.

An Open Data Format: All data in Delta Lake is stored in Apache Parquet format enabling Delta Lake to leverage the efficient compression and encoding schemes that are native to Parquet.

Unified Batch and Streaming: Tables in Delta Lake support batch and streaming interactions. Streaming data ingest, batch historic backfill, and interactive queries work directly.

Schema Management: Delta Lake can enforce defined schemas to ensure that data types are correct and required columns are present, preventing bad data from causing data corruption.

Schema Evolution: Big data is continuously changing. Delta Lake applies changes to table schema automatically, without the need for cumbersome DDL.

Apache Spark Compatibility: Developers can use Delta Lake with their existing data pipelines with minimal change as it is fully compatible with Spark.

cirata

5000 Executive Parkway, Suite 270, San Ramon, CA 94583
US +1 877 (926-3472)   EMEA +44 (0) 114 3039985
APAC +61 2 8211 0620   All other +1 925 380 1728
www.cirata.com