cirata | ORACLE

Cirata Data Migrator for Hadoop to Oracle Lakehouse migrations

Table of contents

Cirata Inc. follows a policy of continuous development and reserves the right to alter, without prior notice, the specifications and descriptions outlined in this document. No part of this document shall be deemed to be part of any contract or warranty.

Cirata Inc. retains the sole proprietary rights to all information contained in this document. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photo copy, recording, or otherwise, without prior written permission of Cirata Inc. or its duly appointed authorized representatives.

Cirata and the Cirata logo are trademarks. All other marks are the property of their respective owners.

Cirata Data Migrator for Hadoop to Oracle Lakehouse migrations

Introduction

A data lakehouse is a modern, open architecture that enables you to store, understand, and analyze all your data. It combines the power and richness of data warehouses with the breadth and flexibility of the most popular open-source data technologies you use today.

The Oracle Lakehouse is built from the ground up on Oracle Cloud Infrastructure (OCI) with the latest AI frameworks and prebuilt AI services. Oracle Lakehouse provides an integrated platform of multiple Oracle cloud services working together with easy movement of data and unified governance, and offers the ability to use the best open-source and commercial tools based on your use cases and preferences (see Figure 1).

Organizations can easily migrate existing or build new open source data lakes in Oracle Lakehouse with fully managed services like Oracle Big Data Service and Oracle Data Flow. Spark, HIVE, Hbase, and many more services can be easily deployed and scaled on OCI.

Oracle Big Data Service provides fully configured, secure, highly available, and dedicated Hadoop and Spark clusters on demand. It offers the commonly used Hadoop components making it simple for enterprises to move workloads to the cloud and ensures compatibility with on-premises solutions.

Oracle Data Flow is a fully managed serverless Spark service that enables you to focus on their Spark workloads with zero infrastructure concepts. It enables rapid application delivery because developers can focus on app development, not infrastructure management.

Many organizations are looking to migrate their on-premises data lakes to leverage the Oracle Lakehouse architecture. However, migrating a data lake from on-premises Hadoop environments to the cloud can be challenging without the right support.

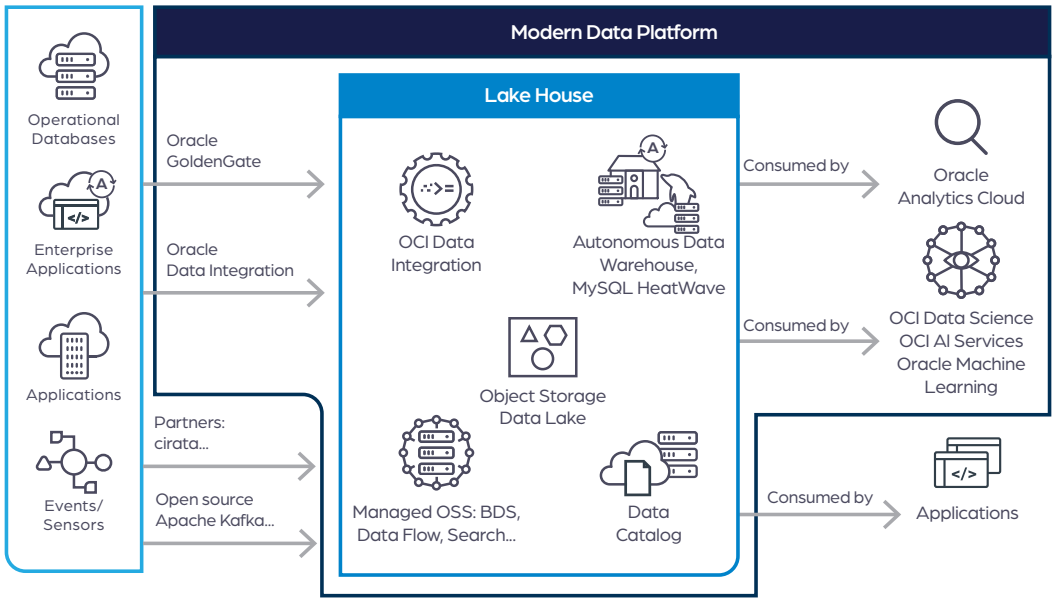

Figure 1: The Oracle Lakehouse

Hadoop migration challenges

Hadoop data migration is hard because of the volume of data and amount of data changes typically occurring in these systems. Traditional data migration approaches relied on tools designed for static data transfer, such as bulk transfer devices or open-source tools like DistCp (Distributed Copy). These require the on-premises systems either to be brought down to prevent data changes from happening during the migration process, or require those responsible for the migration to identify the changes and develop custom solutions to migrate the new and changed data. This adds time and risks to the data migration, and according to industry analysts, results in over 60% of data migration initiatives to go over time, exceed budget, or fail altogether.

Organizations need a migration strategy that reduces business risks, eliminates business disruption, and ensures successful Hadoop to Oracle Lakehouse migrations. For this reason, Oracle has partnered with Cirata to utilize Cirata Data Migrator within its Oracle Cloud Lift Services.

Hadoop migration steps

Typical steps involved in a Hadoop to Cloud migration are as follows (see Figure 2):

- Discovery – Identify the data sets and workloads that are to be migrated to the cloud.
- Planning – Develop a plan and timeline for the phases in which the migration will be performed.
- Data Migration – Perform migration of the required data from the on-premises Hadoop environment to the cloud.
- Workload Migration – Perform migration of the workloads and/or applications from the on-premises environment to the cloud.
- New Analytics Development – Begin to develop new analytics, AI, and machine learning, thereby leveraging the new cloud environment.
- Measure & Act – Perform analytics to measure KPIs, assess performance, make predictions, and enable the business to act appropriately.

To try and simplify their cloud migration, many organizations choose to follow a "lift and shift" migration strategy. This strategy makes the simplistic assumption that the migration can be performed without making any changes to data or the applications. The logic is "just move them as they are to the cloud." This assumption results in many failed projects or projects that exceed their time and costs. It requires either that existing systems be brought down to ensure no data changes occur, or requires that organizations spend time developing custom solutions to handle data changes. Other downsides to this strategy are, first, that it requires organizations to perform a big-bang cut-over of all applications and data at the same time, and second, it doesn't take advantage of new cloud capabilities.

Cirata promotes a data-first approach to data lake migrations. A data-first approach focuses on getting the data moved quickly and not trying to migrate all the existing

Figure 2: Typical Hadoop to Cloud migration steps

applications at the same time. This focus makes the data available to the data scientists faster so they can begin working with the migrated data from day one. This enables for much faster time to new insights and new AI innovations. Organizations can demonstrate a much faster ROI on the cloud migration while the existing on-prem production workloads can continue to run unaffected. This approach also provides flexibility for the application and workload migration. It avoids any big-bang approaches and it provides organizations with time to optimize the workloads for the new cloud environment, ensuring it runs optimally and takes advantage of new capabilities available to them. Organizations can do as much parallel testing as needed to ensure they won't experience any hidden costs, and a data-first approach also gives them time to determine if some of the applications may not need to be migrated at all, but instead replaced with the new development that has been occurring.

Hadoop to Oracle Lakehouse migration with Cirata

Cirata Data Migrator automates the large-scale movement of data and metadata from existing on-premises data lakes, Spark, and Hadoop environments to Oracle Cloud Infrastructure (OCI). Leveraging Cirata's Data capabilities, data migration can occur while the source data is under active change, without requiring any production system downtime or business disruption, and supports complete and continuous data migration.

The reference architecture for using Cirata Data Migrator to automate the data migration to Oracle Lakehouse is depicted in Figure 3.

Data Migrator supports the migration of Hadoop data and Hive metadata from the following sources:

- Cloudera — including CDP (Cloudera Data Platform), CDH (Cloudera Data Hub) or HDP (Hortonworks Data Platform)
- HDFS versions 2.6 and higher

The source systems can be running on Oracle BDA (Big Data Appliance) or custom hardware configurations.

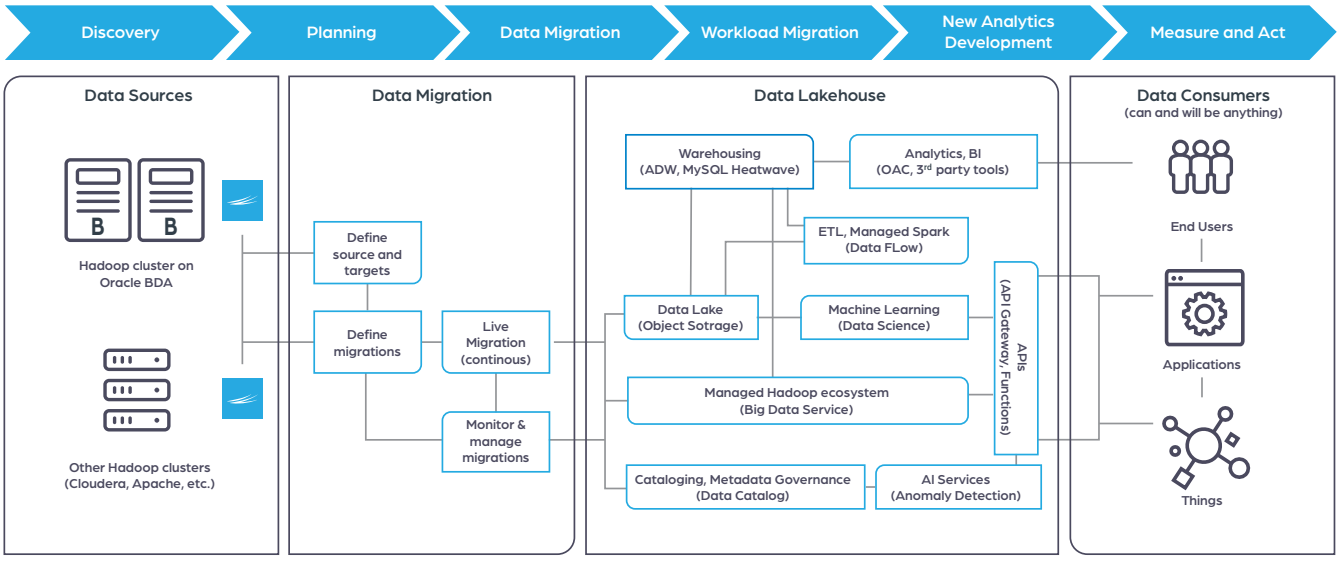

Figure 3: Cirata Data Migrator and Oracle Lakehouse reference architecture

data migrator is deployed on an edge node of the Hadoop cluster. Deployment is performed in minutes with no impact to current production operations. Users can begin to use the product immediately by using the command line, REST API, or user interface (UI) and by performing the following steps:

1 / Define source and targets

During deployment, data migrator automatically discovers the source HDFS cluster so that users only need to define the target environment. This can be done by providing the filesystem type, a display name, the default filesystem path, and some additional configuration settings, as shown in Figure 4.

Note: For Oracle, the filesystem type can either be Oracle Cloud Object Storage or Apache Hadoop if the target is Oracle Big Data Service (BDS), which leverages Oracle's Apache Hadoop distribution.

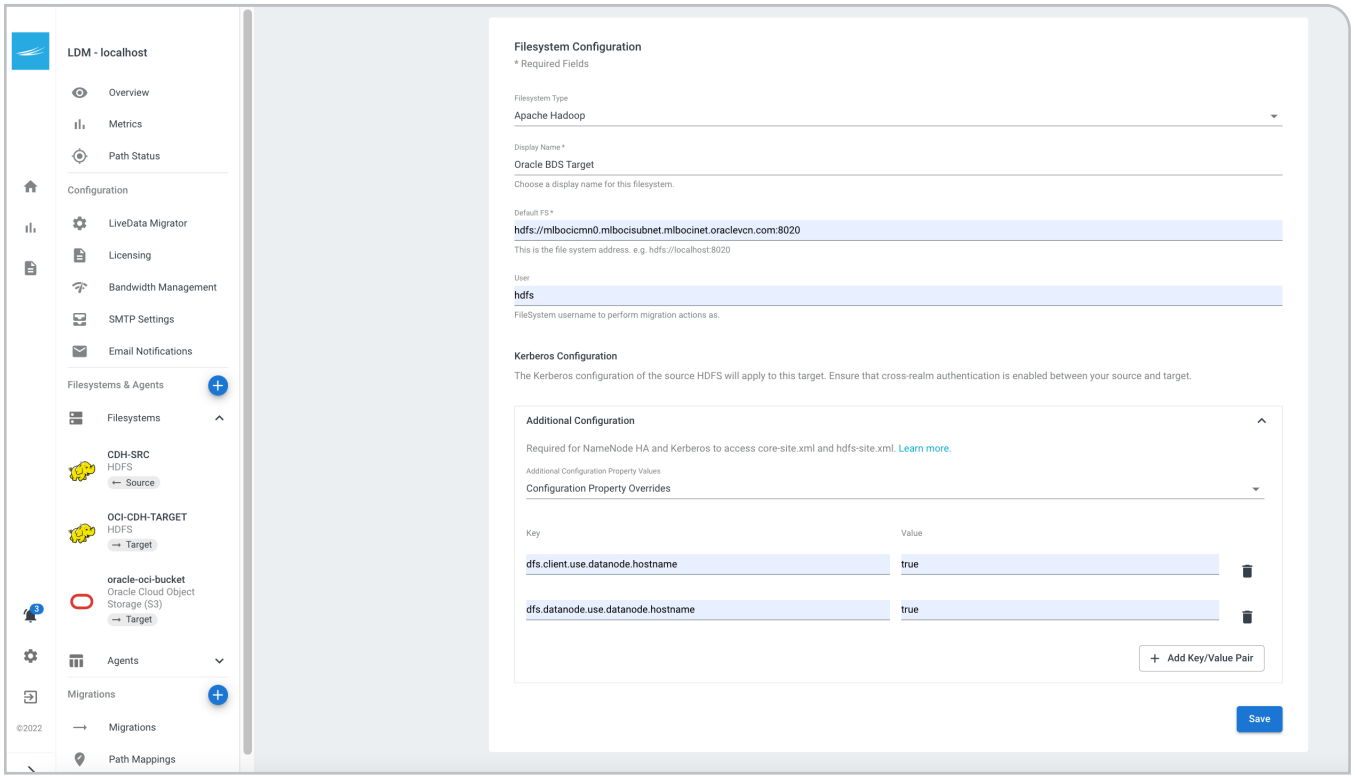

Figure 4: Defining target filesystem using the Cirata UI

2 / Define migrations

Migrations transfer existing data from the source to the defined target. data migrator migrates any changes made to the source data while it is being migrated and ensures that the target is up to date with those changes. It does this while continuing to perform the migration.

Users will typically create multiple migrations so they can select specific content from the source filesystem by path. Users can also migrate to multiple independent filesystems at the same time by defining multiple migration targets.

In order to create a migration, users need to provide a migration name, select the source and target filesystems, and specify the path on the source filesystem to be migrated. Users can optionally apply exclusions to specify rules for data that should be excluded from a migration, and can apply some other optional configuration settings, as shown in Figure 5.

When defining the migrations, users can specify to automatically start the migration and to determine whether it should be a live migration, meaning it will continuously apply any ongoing changes from source to target.

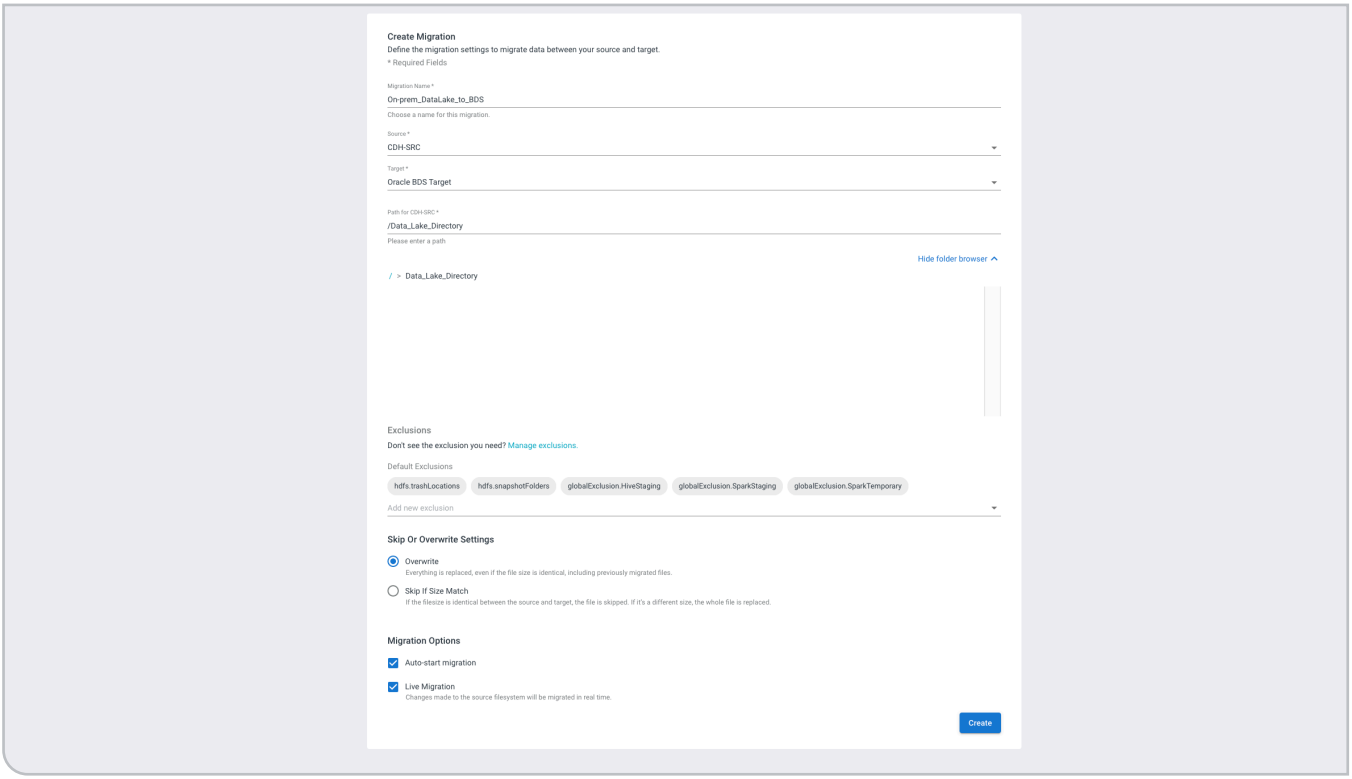

Figure 5: Defining migrations using the Cirata UI

3 / Live migration

If “auto-start” was selected when defining a migration, data will begin to migrate immediately from source to target. Otherwise, a migration will need to be started manually by using the “start migration” option.

Furthermore, if “live migration” was selected when defining the migration, it will run continuously, replicating any changes in real time as they occur from source to target. Otherwise, a one-time migration will be performed.

data migrator also supports migration of Hive metadata from source to target metastores. data migrator connects to metastores through the use of local or remote metadata agents. Metadata rules are then used to define the metadata to be migrated from source to target.

4 / Monitor and manage migrations

Once migrations have been started, they can be monitored by leveraging the Cirata user interface (UI). The UI displays the bandwidth usage for the data being moved, as shown in Figure 6.

Additional migration metrics are available to better understand the migration progress, events yet to be processed, events yet to be migrated, and paths to be scanned.

Existing migrations can be managed via the command line and Cirata UI. Available actions include:

- Assign and remove exclusions from existing migrations
- Start, stop, and resume migrations
- Delete a migration
- Reset a migration to the state it was in before it started
- Monitor failed operations to see date/time, path, and reason for failure

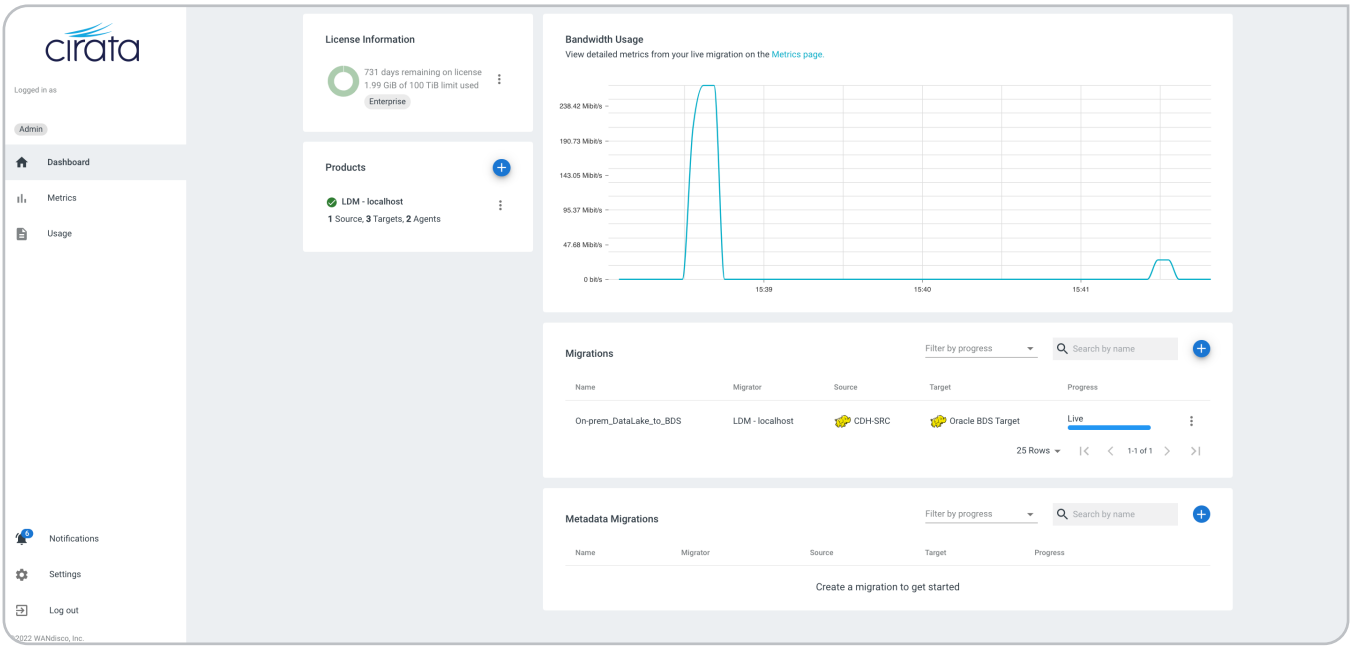

Figure 6: Monitor and manage migrations using the Cirata UI

About Cirata

Welcome to Cirata – a new company with over 45 patents and 15+ years of data science expertise in rapidly integrating high value datasets to leading cloud platforms for game changing AI activation and analytics insights.

We accelerate data-driven revenue growth by automating data transfer and integration to modern cloud analytics and AI platforms without downtime or disruption.

For more information on Cirata, visit www.cirata.com.

ORACLE®

Oracle offers integrated suites of applications plus secure, autonomous infrastructure in the Oracle Cloud. For more information about Oracle (NYSE: ORCL), please visit us at www.oracle.com.

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

wp-dmholm-230913