Security with Cirata Data Migrator

Technical white paper

Security with Cirata Data Migrator

Introduction

Data security is paramount for all organizations. As system, application and user requirements influence how data architectures evolve, supporting platforms must meet their needs and enhance security goals.

Cirata Data Migrator is designed to operate in organizations with stringent security controls and processes. It provides specific features to support those needs, and integrates with enterprise infrastructure to operate securely. This white paper includes information relevant to administrators, security professionals, and compliance officers tasked with maintaining and enhancing data security outcomes at an organizational level. It is structured in three parts:

1. An overview of security requirements that are typical for implementations of Data Migrator,
2. A review of the specific security-related features of Data Migrator, and
3. A description of common patterns of use that apply Data Migrator's security-related features to those requirements.

Enterprise data security requirements

Enterprise data security is important because it helps to protect the most valuable assets of modern organizations: information. Risks related to the generation, aggregation, storage, processing and management of customer data, intellectual property, and financial data are of particular concern to modern enterprises, and require platform capabilities that support security goals. Organizations must consider data security for its contribution to:

- **Data confidentiality:** Breaches of access controls to data can incur business costs for investigation, remediation, legal fees, potential fines and loss of reputation with customers.
- **Market competitiveness:** Customers and users expect and entrust their data to your organization with every interaction. Outcomes that do not meet those expectations harm your competitive position. Business continuity concerns, stakeholder trust and industry best practices are necessary elements to address in maintaining a strong competitive position.
- **Legal compliance:** Many organizations are subject to mandatory legal compliance requirements for data protection, and bound by significant fines.

With increasing scrutiny on data systems and processes, many organizations face the additional need to not just meet their security requirements, but to demonstrate how they meet those needs. Secure data practices may need to encompass the demands of regulations, standards and frameworks such as the EU GDPR, ISO/IEC 27701, the NIST Cybersecurity Framework or similar in order for organizations to meet the needs imposed for conducting business with other entities, or to be competitive against other organizations that can demonstrate their commitment to data security more effectively.

Data security is also an *encompassing* set of requirements, because they can span beyond technical considerations to aspects of financial and process risk, change management and market communication such as security breach disclosures and responses to public vulnerabilities.

Overall, all technologies introduced or used at enterprise scale must provide specific functionality to support an organization's security goals. Because those goals can be broad, and because data integration products such as Cirata Data Migrator can span systems and security boundaries, careful consideration must be given to how they offer features to help meet those requirements.

Security features of Data Migrator

Cirata Data Migrator includes functionality to support enterprise data security, designed in response to the three major types of data security requirements. These are:

1. Technical controls, which help define and limit where data are held, and how access to data is specified and enforced,
2. Preventative security measures, which include monitoring, access control and auditing, and security of data at rest and in transit,
3. Organizational policies, which can define processes and protocols for data handling, along with training and risk management practices.

Technical controls

Migration operation

The core functionality of Data Migrator provides data and metadata transfer between separate, potentially heterogenous environments. Data are transferred as a result of creating an instance of a Data Migration, which is a rule defining:

- the source of data,
- the target of migration,
- the scope of data to be transferred,
- exclusions for specific data assets not to be transferred,
- policies that dictate behavior when pre-existing data reside at the migration target,
- how notifications about changes to source data are acted on,
- the time over which the migration occurs, which can be one-time, with a recurring, scheduled, or on a continuous basis.

Migrations define where and how data is moved. Users have full control over the use of this functionality, which supports security requirements for data locality, availability and disaster recovery.

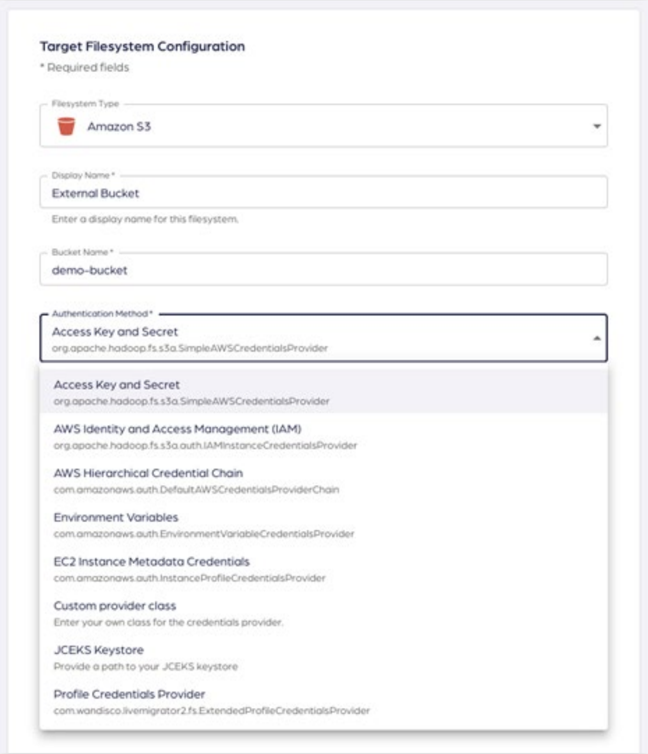
Secrets management

Access to external systems, such as storage platforms, metastores and catalogs is defined using either "file system" or "metadata agent" instances in Data Migrator. Migrations reference file systems and metadata agents to interface with external systems. Users with appropriate authority can create new instances of these to specify how Data Migrator will function. This includes capturing the credentials needed for that interaction, which can also be delegated to an external secrets manager: Hashicorp Vault.

With a secrets store configured with Data Migrator, you can protect specific sensitive filesystem configuration values using references to secrets store keys. You can also protect sensitive application configuration properties by adding key-value pairs to your secrets store and defining a property source in each component's secrets store configuration.

External system authorization

Data Migrator conforms to the specific requirements of each external system with which it can integrate, and honors their needs for authorization. The specific mechanism will depend on the particular system, and Data Migrator defaults to the most secure option for providing credentials in every instance. As an example, when referencing an Amazon S3 bucket, users can select among a variety of options for authorization:



Target Filesystem Configuration

* Required fields

Filesystem Type: Amazon S3

Display Name *: External Bucket

Enter a display name for this filesystem.

Bucket Name *: demo-bucket

Authentication Method *: Access Key and Secret
org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider

Access Key and Secret
org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider

AWS Identity and Access Management (IAM)
org.apache.hadoop.fs.s3a.auth.IAMInstanceCredentialsProvider

AWS Hierarchical Credential Chain
com.amazonaws.auth.DefaultAWSCredentialsProviderChain

Environment Variables
com.amazonaws.auth.EnvironmentVariableCredentialsProvider

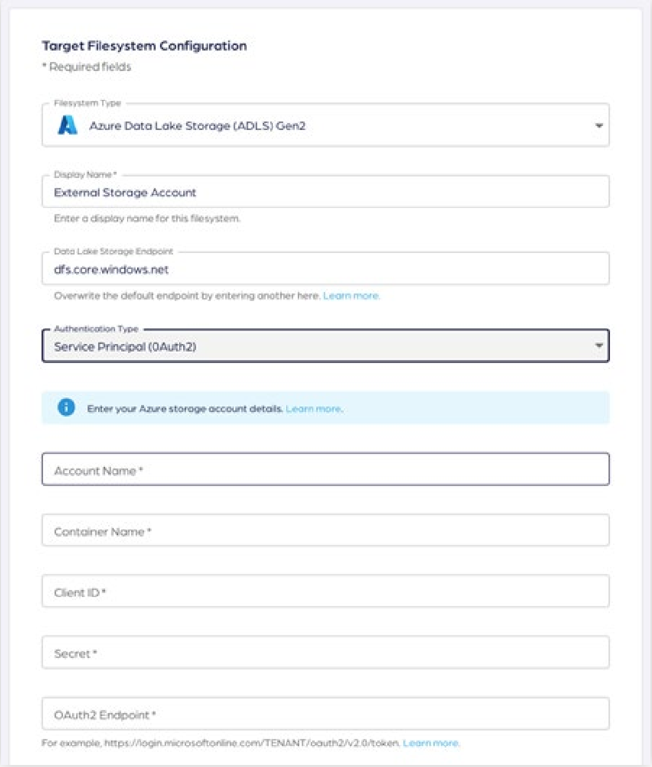
EC2 Instance Metadata Credentials
com.amazonaws.auth.InstanceProfileCredentialsProvider

Custom provider class
Enter your own class for the credentials provider.

JCEKS Keystore
Provide a path to your JCEKS keystore

Profile Credentials Provider
com.wandisco.livemigrator2.fs.ExtendedProfileCredentialsProvider

While for Azure Data Lake Storage Gen 2, Data Migrator allows either Service Principal (OAuth 2) or Shared Key authentication:



Target Filesystem Configuration

* Required fields

Filesystem Type: Azure Data Lake Storage (ADLS) Gen2

Display Name *: External Storage Account

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: Service Principal (OAuth2)

Enter your Azure storage account details. Learn more.

Account Name *

Container Name *

Client ID *

Secret *

OAuth2 Endpoint *

For example, https://login.microsoftonline.com/TENANT/oauth2/v2.0/token. Learn more.

Preventative security measures

Access controls

Data Migrator uses a standard and comprehensive approach to role-based access control that integrates with an organization's existing directory infrastructure to define user and group roles and their memberships. Cirata Data Migrator then applies role definitions to enforce access control to functionality throughout its operation, including for the user interface, command-line tooling and REST API.

UI, API and CLI security

Once configured to use a directory , Data Migrator automatically enforces access controls through the user interface. Users will have either no access, read-only access, migration manager access or administrator access to the product based on their role allocation driven by groups defined in the external directory. Data Migrator supports mapping directory roles to administrator, migration manager, and read-only roles in the product, providing comprehensive allocation to common functionality.

Access to the REST API exposed by Data Migrator and the command-line interface is also restricted by assigning users to permission-based roles. Configure a filter to find groups within your directory tree, then map one or more directory groups to each role to apply that role to the users in those groups.

By enforcing common access controls based solely on role allocation, managing user access is straightforward, and can be done through existing directory tooling independently of Cirata's product.

Management groups

Data Migrator also supports the allocation of users to management groups, which grant the authority to manage individual data or metadata migrations to group members. This allows a federated security model, where subsets of data and metadata migrations can be managed independently. Management groups augment the role-based access controls in the product to allow a more flexible allocation of privileges based on your data management needs, rather than just your organization structure.

Monitoring

Data Migrator collects audit information throughout operation, including:

- Audit logs with events that are related to or have been actions by users in the UI, such as when they logged in or when they created a migration,
- Audit logs based on the operation of data and metadata migrations, capturing every activity performed against a target in transferring data or metadata,
- Data Migrator activities, which include information about events that are related to or have been actions by users in a Data Migrator instance, such as when they configured a filesystem or when they created a data migration, and
- Hive Migrator activities, which includes information from the logs about events that are related to or have been actions by users in a Hive Migrator instance, such as when they configured an agent or when they created a metadata migration.

By retaining audit records of all operation activities, Data Migrator provides advantages in environments that need records of actions performed by users or the systems that they affect using the platform.

Security of data in transit

Much of Data Migrator's runtime operation involves network communication with external systems. Besides adhering to and supporting their individual authentication requirements, Data Migrator will default to secure network communication, typically using TLS for encryption where it is offered by the external system.

Security of data at rest

Data Migrator does not store user data, but orchestrates and manages how it is transferred between external systems. The security of data at rest is defined by the capabilities of those external systems, and Data Migrator allows users to take advantage of that functionality by choosing which instances of the various data storage boundaries are employed.

Organizational policies

By its very nature as a commercially-supported product, Data Migrator offers advantages beyond open-source or DIY approaches to data integration needs that include the types of organizational policies directed to enterprise security. Data protection strategies that include backups, disaster recovery processes and plans, data handling guidelines, etc. can benefit from the expertise gained by Cirata from previous production implementation of Data Migrator.

Cirata also works with both OEM and services partners to provide and implement the Data Migrator technology, further extending the expertise available to customers, particularly in highly regulated industries such as financial services.

Implementations of data security policies must take into account the features and capabilities of all systems involved. Regular training, process review, penetration testing and exercising of disaster scenarios must accompany day-to-day activities to ensure that security policies are being met effectively.

As a commercial product, Data Migrator includes:

- a support model that defines specific terms and conditions that include maintenance and support with defined priority levels and response times, root cause analysis for information purposes, trouble ticket reporting systems, product fixes, etc.
- a software bill of materials that includes full information about included third-party libraries and versions,
- comprehensive product documentation,
- customer support services for implementation design, review, upgrades, enhancements, etc.
- resource libraries,
- a knowledge base, and
- community forums.

Common Data Migrator practices for security

Through regular application of Data Migrator with our customers, Cirata has gained significant exposure to good and bad security practices across a broad range of enterprises. Common patterns of use that apply Data Migrator's security-related features to security requirements emerge from that work, and are captured by our customer success team as part of our ongoing services improvements.

From that work a list of the top 5 best practices that are directly possible with Cirata's technology have been identified, and are representative of the benefits of using Data Migrator in place of ad-hoc approaches to large-scale data integration requirements. These are:

1. Integration with role-based controls at an organization level.
2. Federation of responsibilities for data transfer to data owners.
3. Operational management of data transfer infrastructure by centralized teams.
4. Consistent failure exercising.
5. Regular audit and review of data security processes.

Integration with RBAC

The most effective means of ensuring compliance with organizational roles is the integration of Data Migrator with a central directory from which roles can be mapped to Data Migrator functionality. This allows security controls to be enforced independently of Cirata's product in a consistent way across all elements of the organization.

1. Use a central directory.
2. Maintain suitable groups and roles.
3. Audit group membership and role scope.

Federation of data transfer responsibilities

The most direct users of data should be responsible for the management of data transfer to the systems from which they will access it, rather than relying on a central team with this responsibility. This allows user-specific security controls to be enforced, rather than providing a single group broader access to data assets than would otherwise be required if data transfer needs were not in place.

- Delegate migration management to users.
- Enforce controls through Data Migrator's functionality.
- Maintain audit records of user activity for regular review.

Manage data transfer infrastructure in a central team

Unlike the federated management of data transfers themselves, Cirata recommends that a centralized team is given the responsibility of operating the infrastructure by which data transfer requirements are met. This is because there are benefits to having central visibility and control of limited resources (cluster and storage capacity, bandwidth, compute overheads, etc.) that cannot be readily delegated because of shared access to common resources.

- Centralize infrastructure operations.
- Maintain delegation of migration management so that it is not performed by infrastructure operations.
- Conduct regular reviews of system capacity and low-level performance and tuning.

Exercise failure scenarios consistently

The most robust operational approach to data security includes consistently-applied exercises that identify and analyze the organization's response to data security breaches. By performing this work in a structured rather than ad-hoc way, the organization can measure its performance against well-defined criteria and improve over time. Ad-hoc approaches are also useful, but do not answer the potential degradation of otherwise healthy processes and systems.

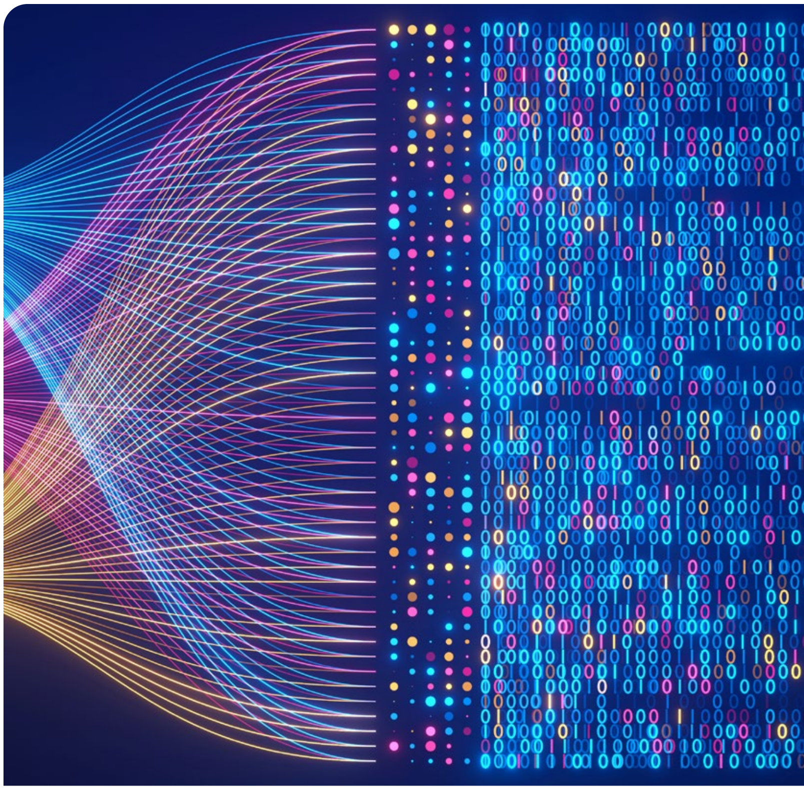
- Augment ad-hoc security scanning and testing with consistently-applied exercises.
- Analyze their outcomes and document improvements.
- Measure the impact of process and technology changes to security goals over time.

Regular audit and review of processes

While it is represented in each of the other top five best practices, the significance of regular audit and review of process is sufficient to stand on its own as the final one in this set. You can only improve what you can measure, and measurement requires auditable records of activities, outcomes and change to be effective.

About Cirata

Trusted by global brands and industry leaders for more than 15 years, Cirata specializes in the migration of Hadoop data lakes into leading cloud platforms to enable game-changing Artificial Intelligence ("AI") and analytics. With Cirata, data leaders can leverage the power of AI and analytics across their entire enterprise data estate to freely choose analytics technologies, avoid vendor, platform, or cloud lock-in while making AI and analytics faster, cheaper, and more flexible. Cirata's portfolio of products and technology solutions make strategic adoption of modern data analytics efficient, automated, and risk-free. In addition, leveraging our patented technologies, including the Distributed Coordination Engine ("DConE®"), our DevOps solutions integrate effortlessly with your existing source code management to increase security, minimize risk, reduce latency, and improve collaboration across globally distributed development teams. For more information visit www.cirata.com.



5000 Executive Parkway, Suite 270, San Ramon, CA 94583
US +1 925 380 1728
EMEA +44 (0) 114 3039985
APAC +61 2 8211 0620
All other +1 925 380 1728

www.cirata.com