PharmaSUG 2022 - SI-097

Ensuring Distributed Data Custody on Cloud Platforms

Ben Bocchicchio, SAS Institute, Cary, North Carolina, USA
Sandeep Juneja, Biogen, Cary, North Carolina, USA

Introduction

Many companies as part of their cloud strategies have moved significant amounts of data into multiple cloud vendors' storage. Each cloud platform has its own Identify and Access Management process. Generating a holistic view of the data spread across multiple cloud platforms generates new challenges.

In this paper, different options will be discussed as how to establish secure connections through various authentication mechanisms used between various cloud platforms. The objective is to establish a controlled and audited way to generate data in the cloud while maintaining data custody and data integrity across various cloud providers.

The need to trace data can be illustrated by the following use cases:

- Data stored in one cloud solution needs to be merged with data stored in another cloud solution
- Analyses completed in one cloud solution need to be posted to another cloud solution

How can the movement of data be traced for repeatability and consistency? Audit Trails can be used to track data exchanges between cloud solutions. Through these audit trails, users can ensure the proper chain of custody. Since there are too many cloud providers to be examined within this review, I have chosen three: AWS, Azure, and a SAS cloud solution (Life Science Analytics Framework).

Discussion

The generation of accurate audit trails generated within a specific cloud application vary greatly. Details below outline what is needed for each system, specifically what Authentication method is best to use and how to configure the auditing

AWS - Authentication

AWS supports two types of authentications: Identify and Access management (IAM) and federated sign-in through AD and ADFS. While IAM is easy to setup, it allows the user to set a local name which could vary across cloud systems. Having different user information across environments will make it difficult to ensure consistency across cloud systems. It is recommended to use federated sign-in through AD and ADFS. This streamlines identity management by sourcing and managing all your user identities from one Active Directory source.

AWS - Audit Trail

By default, AWS has S3 access logs, however these audits are not sufficient to ensure traceability. It is recommended to use Object Level Logging in CloudTrail. CloudTrail has central auditing and logging and the ability to control what buckets, prefixes, and objects will be audited, as well as what types of actions to audit.

Azure – Authorization

Azure supports four authentication methods: shared key authorization, role-based access control (RBAC), shared access signature and Access Control Lists (ACL). How are permissions evaluated? - During security principal-based authorization, permissions are evaluated first by Role Assignment and then by ACLs.

Storage account access keys (shared) are a simpler method to access data from the Azure Storage Account, however, audit trails are obscured due to this being a 'generic' account access. It is recommended to use Azure AD user accounts (RBAC). When using this Authorization method, users have identities in a centralized Identity and Access Management system and can use their own identity to access the data in the Storage Account hence accurate audit trail can be captured.

Azure - Audit Trails

By default, Azure Monitor Logs consolidates logs and metrics from multiple services and other data sources. What is needed is to create a diagnostic setting to collect platform metrics, activity logs, and resource logs into a Log Analytics workspace in Azure Monitor. Log Analytics allows you perform advanced analysis of log data using a fully featured query language. This is how to extract the required log information to support traceability.

SAS® Life Science Analytics Framework (LSAF)

By default, LSAF has an internal Active Directory that supports the creation of User IDs and Passwords for use within the system. LSAF also supports SSO integration from a cloud authorization method. Using this capability allows for capturing a consistent User ID across cloud systems.

SAS Life Science Analytics Framework (LSAF) - Audit Trails

LSAF provides an audit trail that tracks all editable actions within the application. The audit trail is programmatically accessible through SAS code.

Conclusion:

Data traceability across cloud environments is accomplished by 1) using consistent User IDs and 2) capturing consistent audit trails records. The establishment of these varies across cloud providers. Provided this standardization has been implemented, the audit trails can then be extracted and combined to prove data traceability across cloud providers. This is critical for both internal and external auditors looking to trace the source data used for analysis and reporting.

Useful Links

AWS

- AWS Federated Authentication with Active Directory Federation Services (AD FS) https://aws.amazon.com/blogs/security/aws-federated-authentication-with-active-directory-federation-services-ad-fs/
- S3 Permissions Overview: S3 Permissions Classification https://shunliz.gitbooks.io/aws-certificate-notes/content/s3/s3-permissions-overview.html

Azure

- Guidance for using Azure Storage Explorer with Azure AD authorization for Azure Storage Data Access https://medium.com/microsoftazure/guidance-for-using-azure-storage-explorer-with-azure-ad-authorization-for-azure-storage-data-access-663c2c88efb
- Tutorial: Monitor Azure resources with Azure Monitor https://docs.microsoft.com/en-us/azure/azure-monitor/essentials/monitor-azure-resource

SAS

- SAS® Life Science Analytics Framework https://www.sas.com/en_us/software/life-science-analytics-framework.html

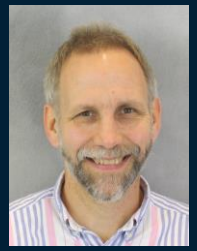
Ensuring Distributed Data custody on Cloud Platforms SI-097

Ben Bocchicchio, SAS Institute, Cary, North Carolina, USA

Sandeep Juneja, Biogen, Cary, North Carolina, USA

Ensuring Distributed Data custody on Cloud Platforms

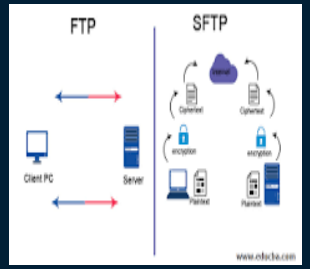
Biography



Ben is a Senior Consulting Manager at SAS and has more than 25 year of experience in designing and implementing solutions for various Health Care and Life Science clients.

Sandeep is AWS & Azure certified Solutions Architect. He is Associate Director of Data Science at Biogen and has more than 18 years of experience in designing and implementing Cloud based solutions for various Health Care and Life Science clients.

Where is your clinical data?



Ensuring Distributed Data custody on Cloud Platforms

Introduction

- Governance is the oversight role and process by which companies manage and mitigate business risk
- Compliance ensures that an organization has the process and internal controls to meet the requirements imposed by governance body
- Data Custody
  - Authentication
  - Authorization (Permissions)
  - Audit Trail

Ensuring Distributed Data custody on Cloud Platforms

Usecase(s)

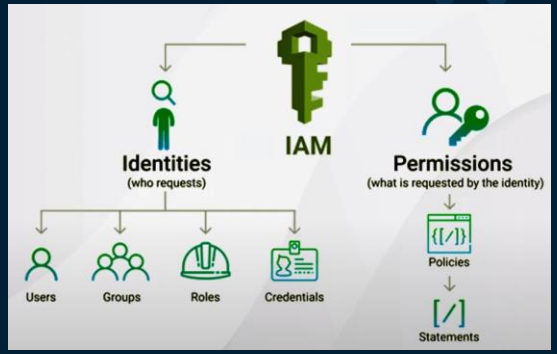
- Data stored in one cloud solution needs to be merged with data stored in another cloud solution
- Analyses completed in one cloud solution needs to be posted to another cloud solution
- Audit Trails tracking data exchange between cloud solutions are required to maintain chain of custody

Ensuring Distributed Data custody on Cloud Platforms

AWS - Authentication

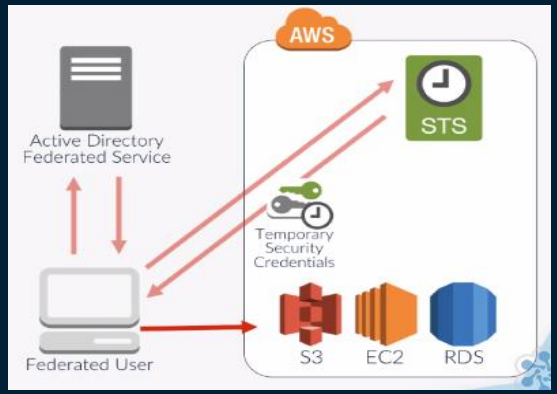
- AWS Identity and Access Management (IAM)

  Allows to centrally manage users, security credentials such as access keys, and permissions that control which resources users can access



- Federated sign-in through AD and ADFS

streamline identity management by sourcing and managing all your user identities from Active Directory

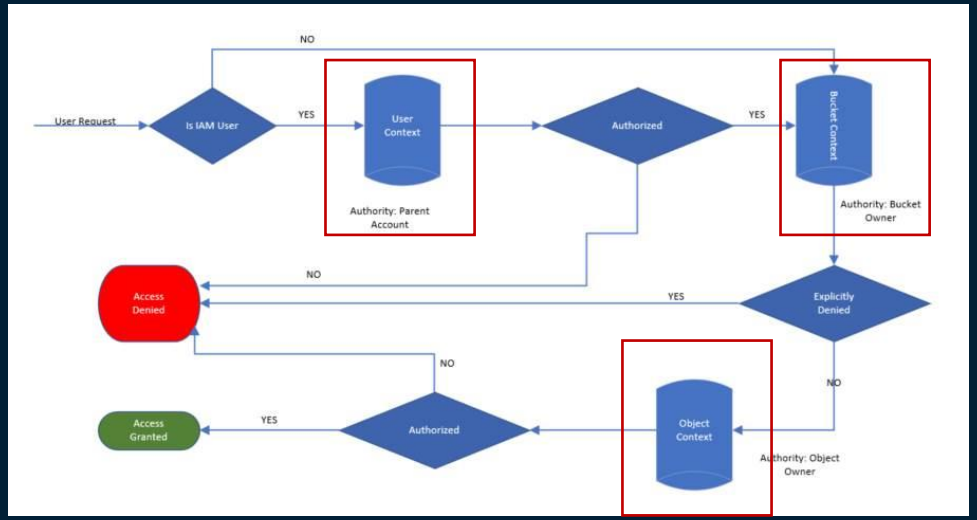


Reference: https://aws.amazon.com/blogs/security/aws-federated-authentication-with-active-directory-federation-services-ad-fs/

Ensuring Distributed Data custody on Cloud Platforms

AWS - Authorization

- S3 evaluates the policies in 3 context to authorize or deny request
  - User context – access policies attached to user
  - Bucket context - access policies owned by the bucket owner
  - Object context - access policies owned by the Object owner

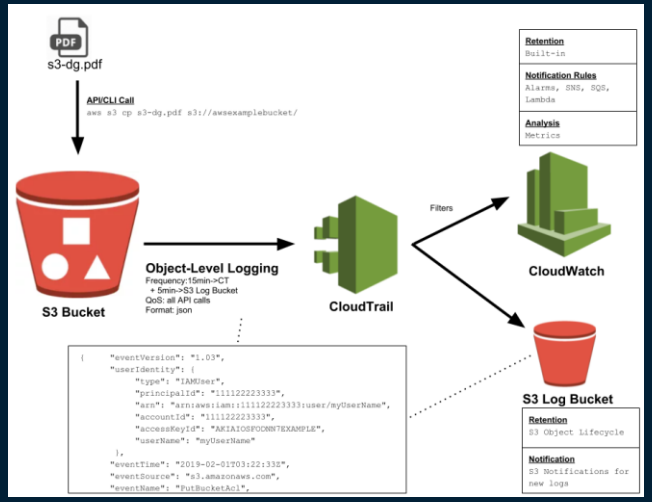


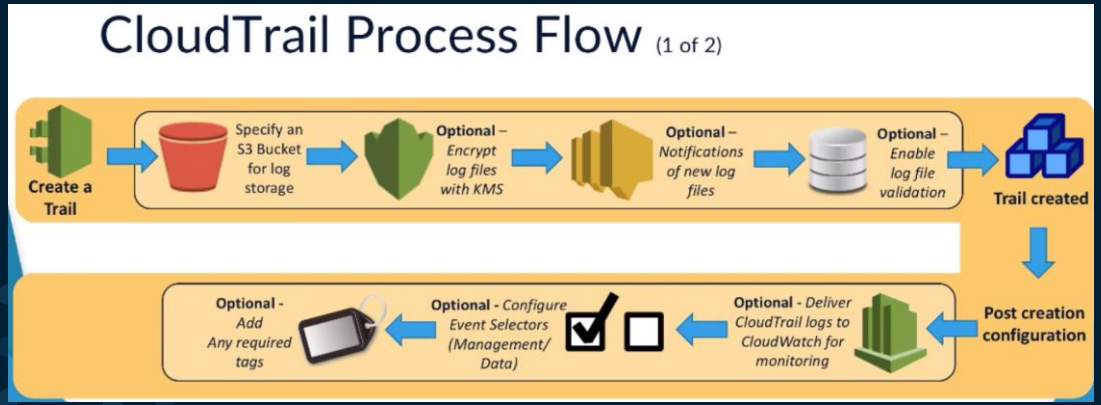
Reference : https://shunliz.gitbooks.io/aws-certificate-notes/content/s3/s3-permissions-overview.html

Ensuring Distributed Data custody on Cloud Platforms

AWS - Audit Trail

- Object Level Logging
  - central auditing and logging in CloudTrail
  - Ability to control what buckets, prefixes, and objects will be audited, and what types of actions to audit.



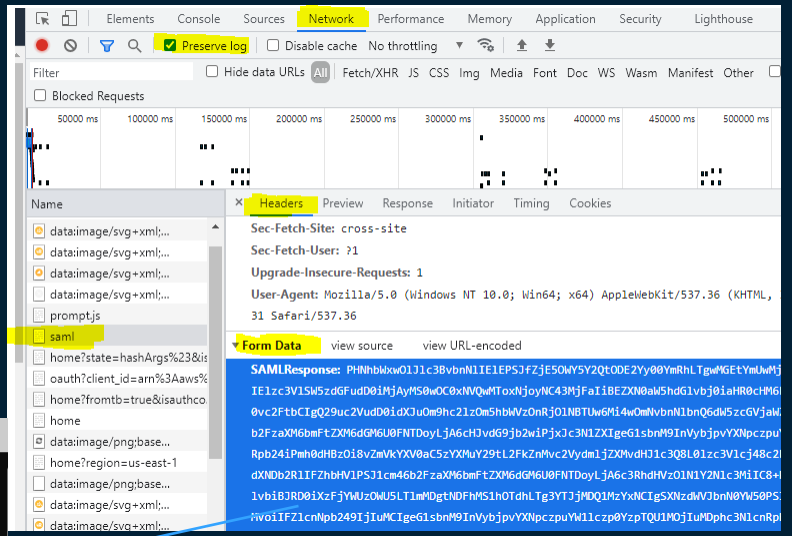


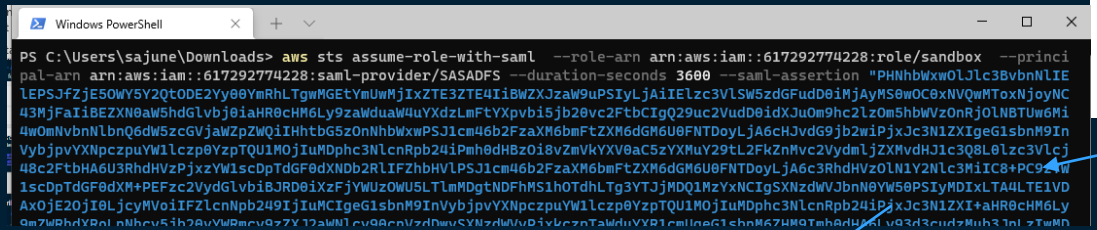
It's recommended to use AWS CloudTrail data events instead of Amazon S3 access logs. CloudTrail data events contain more information

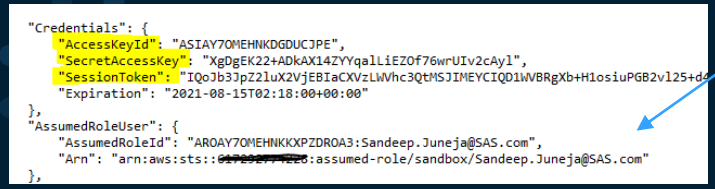
Ensuring Distributed Data custody on Cloud Platforms

AWS – Mount S3 as local drive

- Go to SSO website and login - https://<fedauth server>.com/adfs/ls/IdpInitiatedSignon.aspx
- For e.g. – https://fedauth.sas.com/adfs/ls/IdpInitiatedSignon.aspx
- Reference you tube video – https://www.youtube.com/watch?v=ACWRmFjG9V8



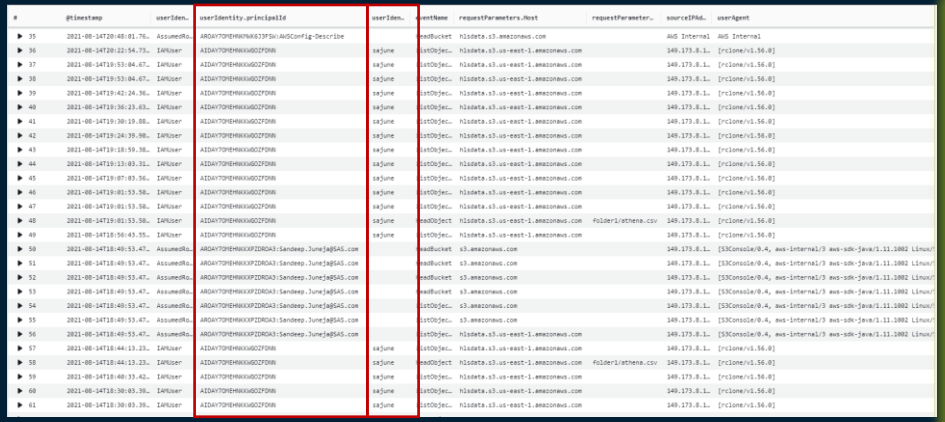




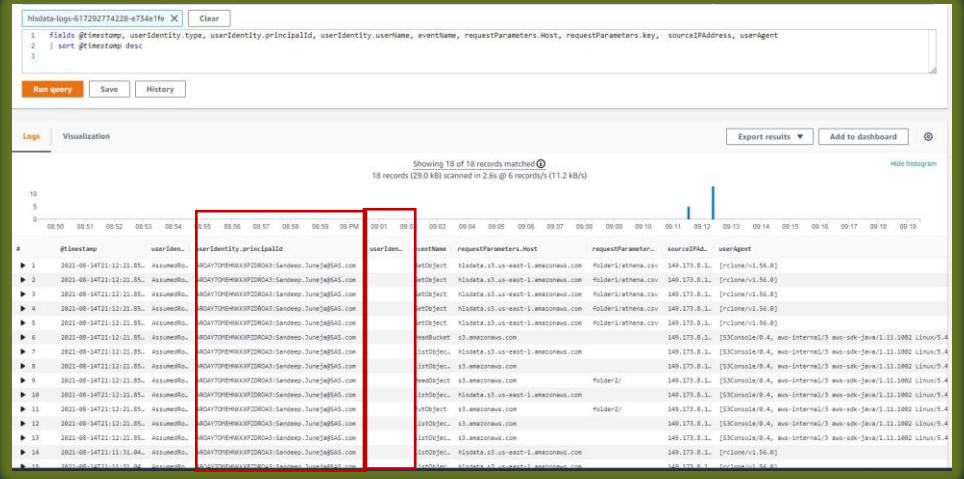
Ensuring Distributed Data custody on Cloud Platforms

AWS Audit Trail

- Using IAM user to mount drive
  - Obscured Audit Trail



- Using AD User to mount drive
  - Concise Audit Trail

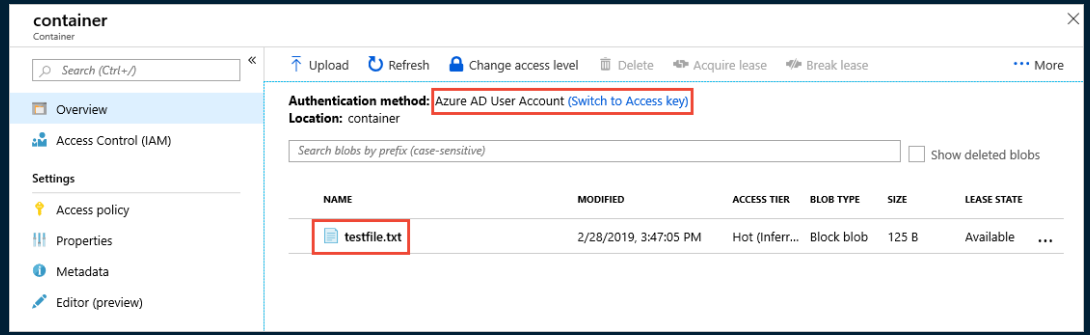


Ensuring Distributed Data custody on Cloud Platforms

Azure Authorization

- Storage account access key - simpler method where an Access key can be used to access data from the Storage Account, however audit trail is obscured (generic account access)

- Azure AD Account - Users have identities in a centralized Identity and Access Management system and use their own identity to access the data in the Storage Account hence accurate audit trail can be captured



Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/authorize-data-operations-portal

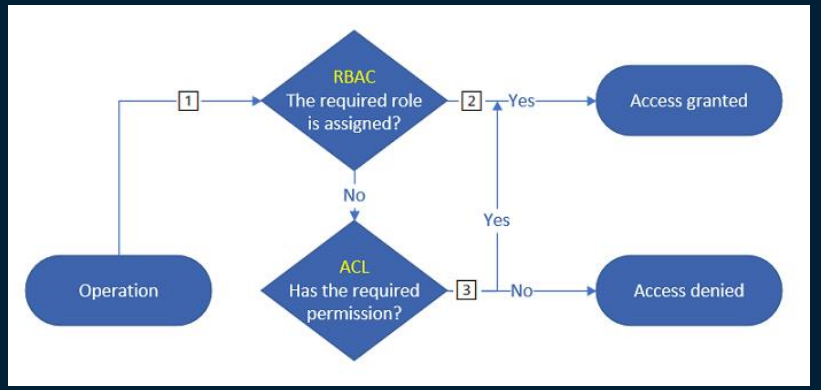
Ensuring Distributed Data custody on Cloud Platforms

Azure Authorization

- Azure Storage (Data Lake Storage Gen2) supports the following authorization mechanisms
  - Shared Key authorization
  - Role-based access control (Azure RBAC)
  - Shared access signature (SAS) authorization
  - Access control lists (ACL)

How permissions are evaluated - During security principal-based authorization, permissions are evaluated in the following order.

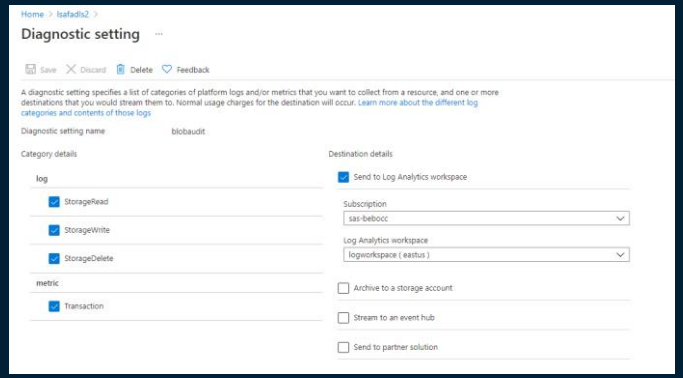
1. Role Assignments
2. ACLs

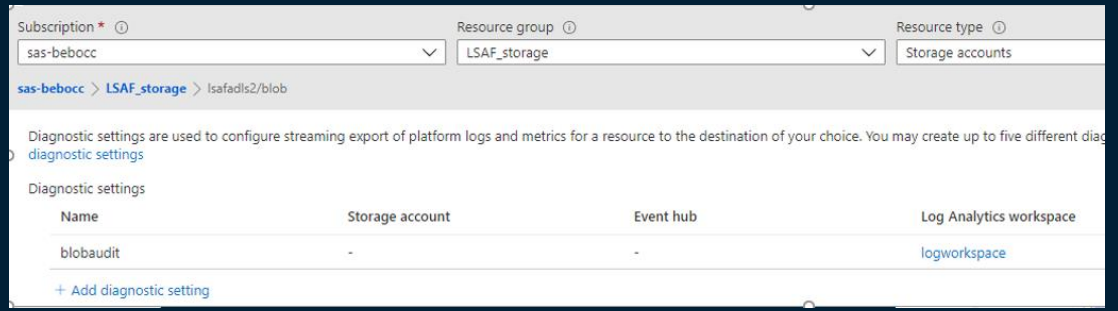


Ensuring Distributed Data custody on Cloud Platforms

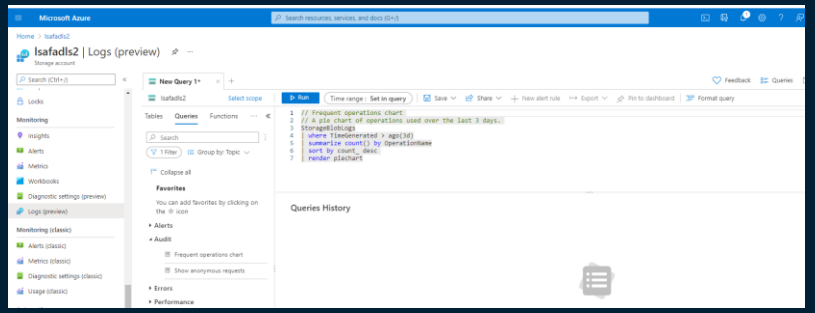
Azure Audit Trail

- Azure Monitor Logs - consolidates logs and metrics from multiple services and other data sources
  - create a Diagnostic setting to collect platform metrics, activity log, and resource logs into a Log Analytics workspace in Azure Monitor.





- Log Analytics allows you perform advanced analysis of log data using a fully featured query language

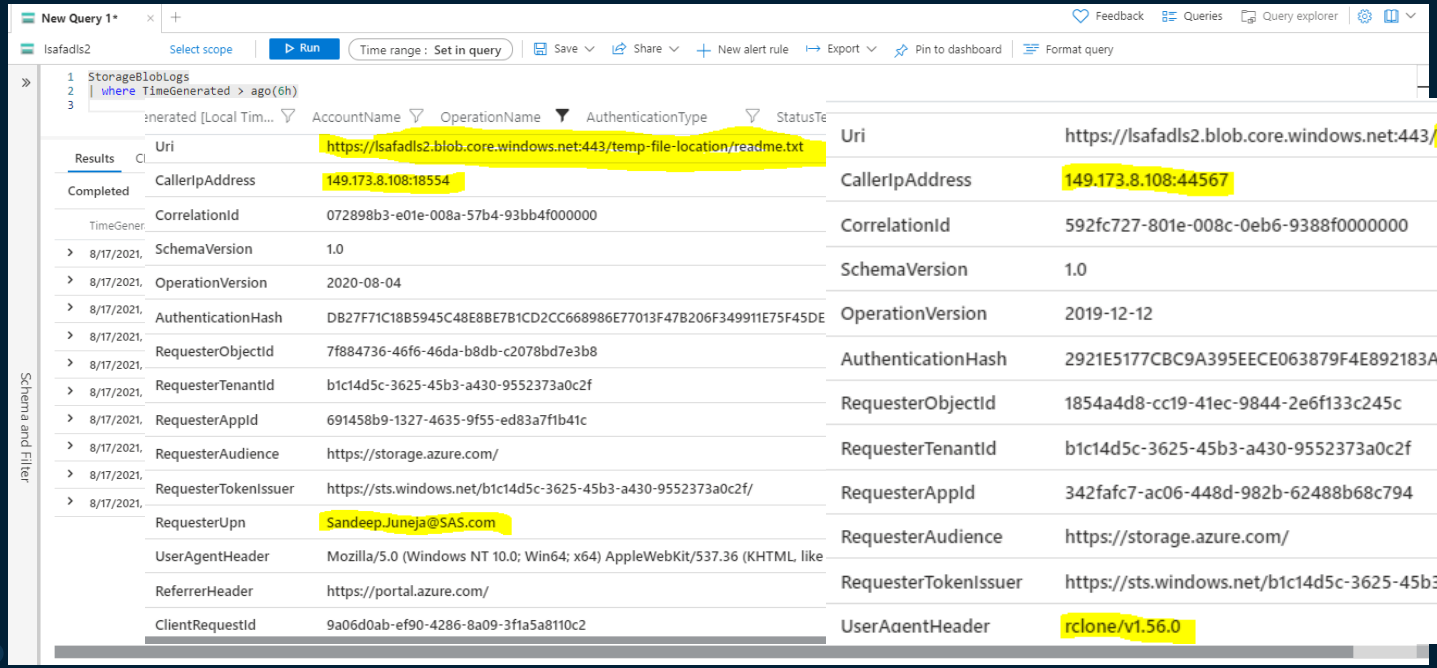


Reference: https://docs.microsoft.com/en-us/azure/azure-monitor/essentials/monitor-azure-resource

Ensuring Distributed Data custody on Cloud Platforms

Azure Audit Trail

- Audit Records

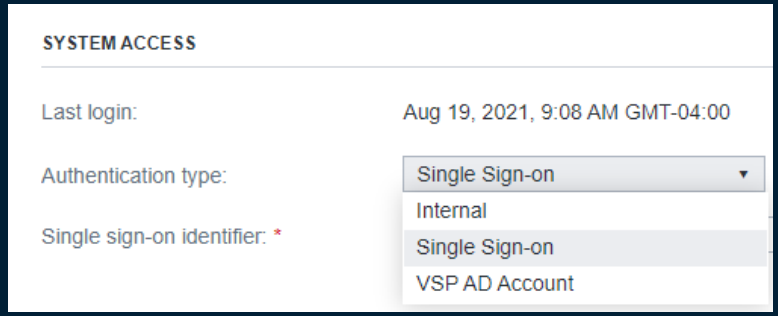


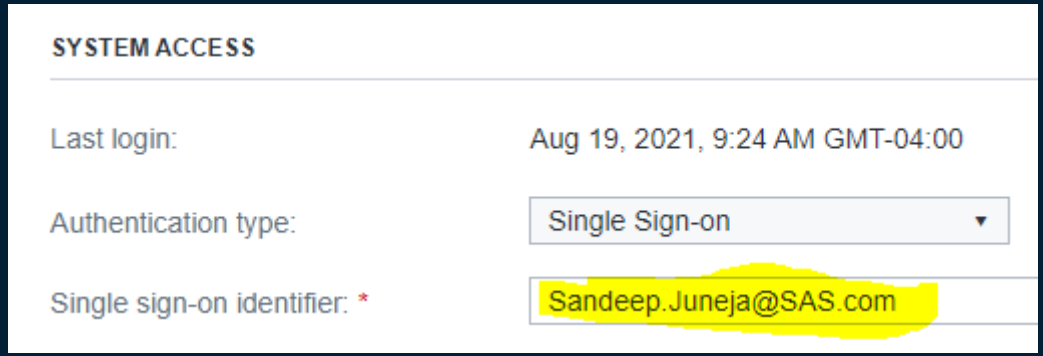
Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/authorize-data-operations-portal

Ensuring Distributed Data custody on Cloud Platforms

SAS Life Science Analytics Framework (LSAF) Authentication

- Authentication
  - Internal
    - Authentication mechanism controlled by application
  - Active Directory
    - Authentication passed o AD
  - Single Sign-on
    - SAML

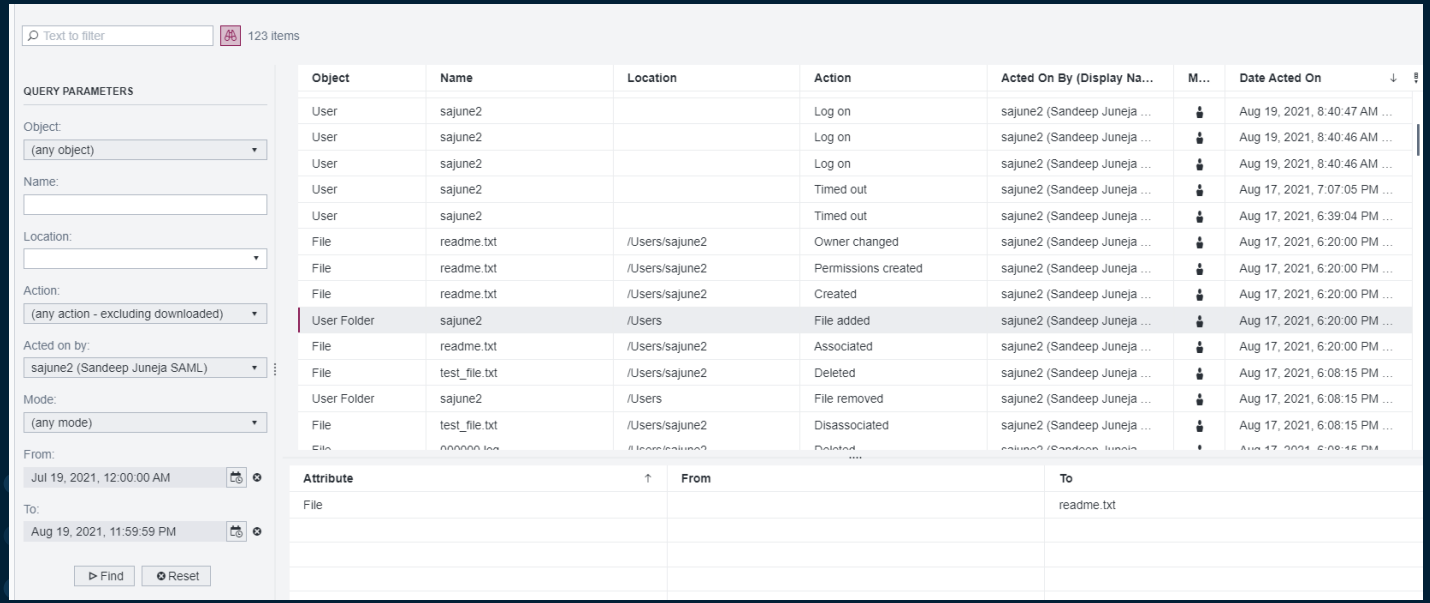




Ensuring Distributed Data custody on Cloud Platforms

SAS Life Science Analytics Framework (LSAF) Audit Trail

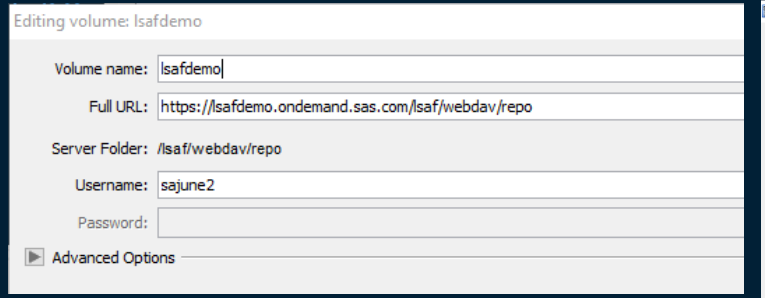
- Audit Trail – Tracks all editable actions

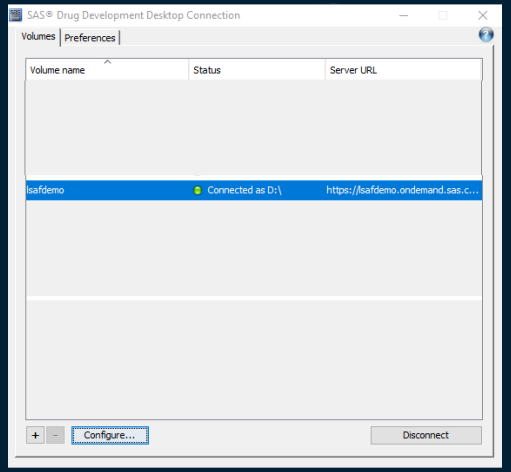


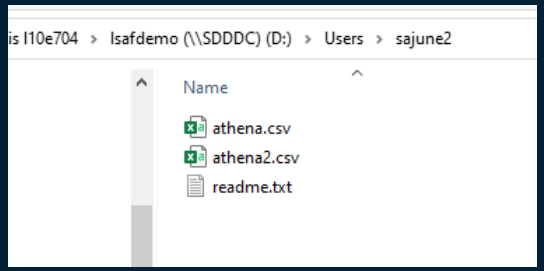
Ensuring Distributed Data custody on Cloud Platforms

Mounting SAS Life Science Analytics Framework (LSAF) as local drive

- Using SAS Drug Development Desktop Connect
- Configure URL and credentials



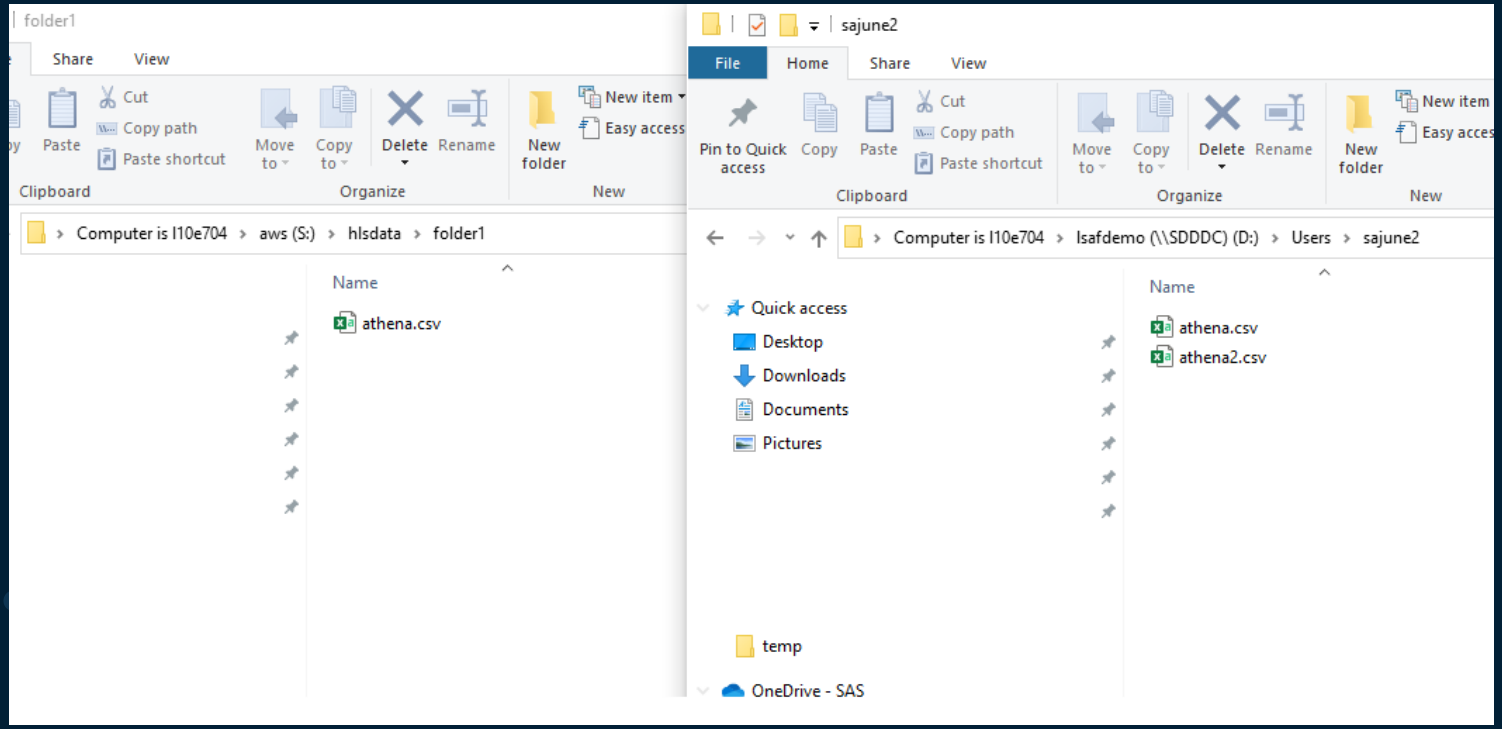




Ensuring Distributed Data custody on Cloud Platforms

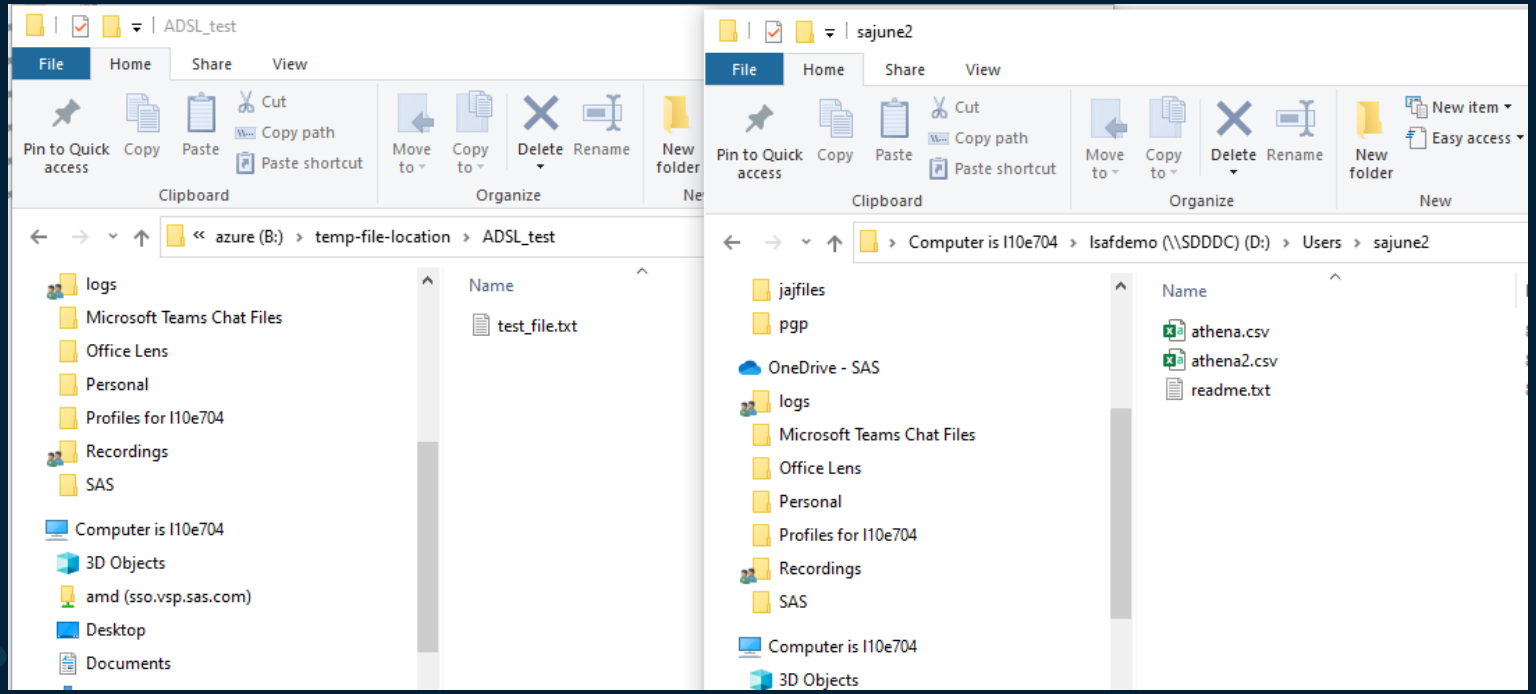
AWS S3 To LSAF Repository

- Data Exchange between cloud storages using drive mounts



Ensuring Distributed Data custody on Cloud Platforms

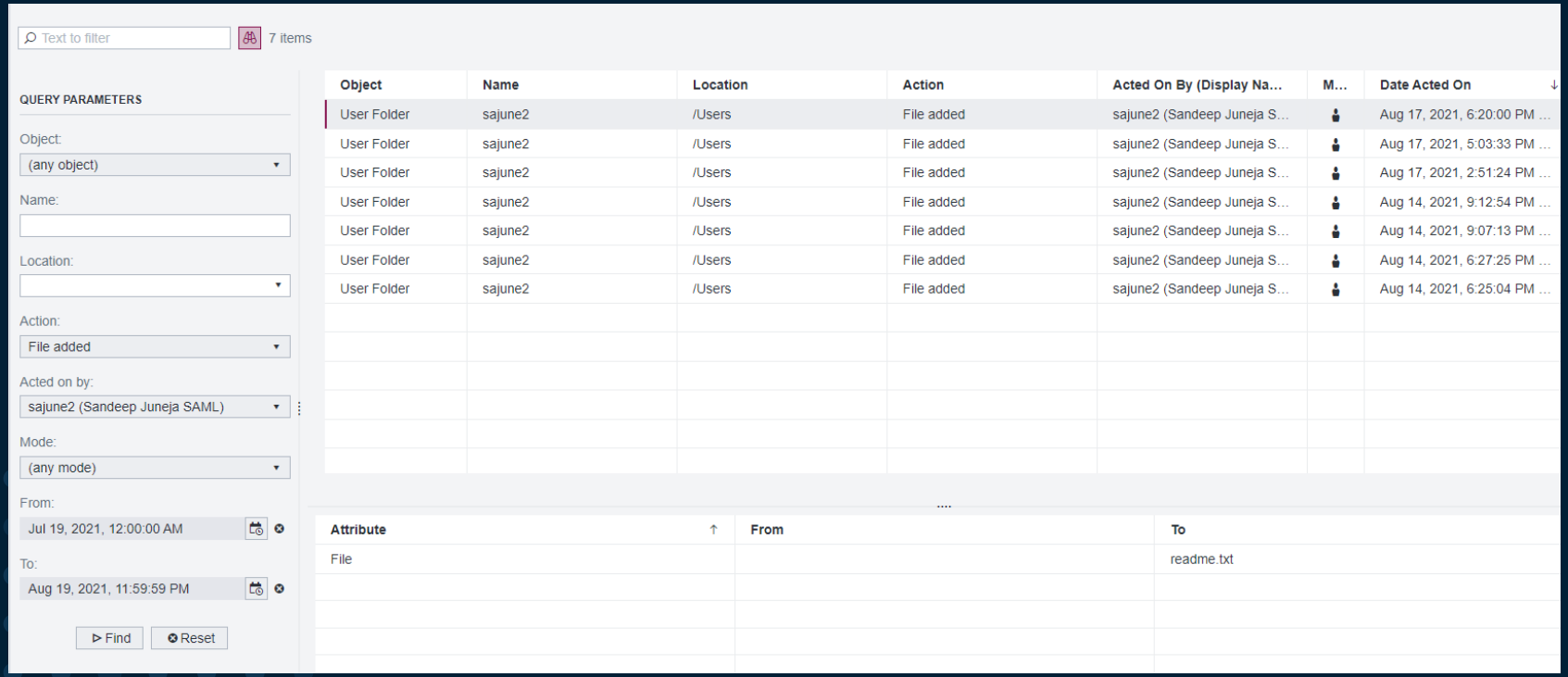
Azure Storage account to LSAF Repository



Ensuring Distributed Data custody on Cloud Platforms

LSAF Audit Records

- Audit Records



Ensuring Distributed Data custody on Cloud Platforms

Data Traceability across cloud Environments

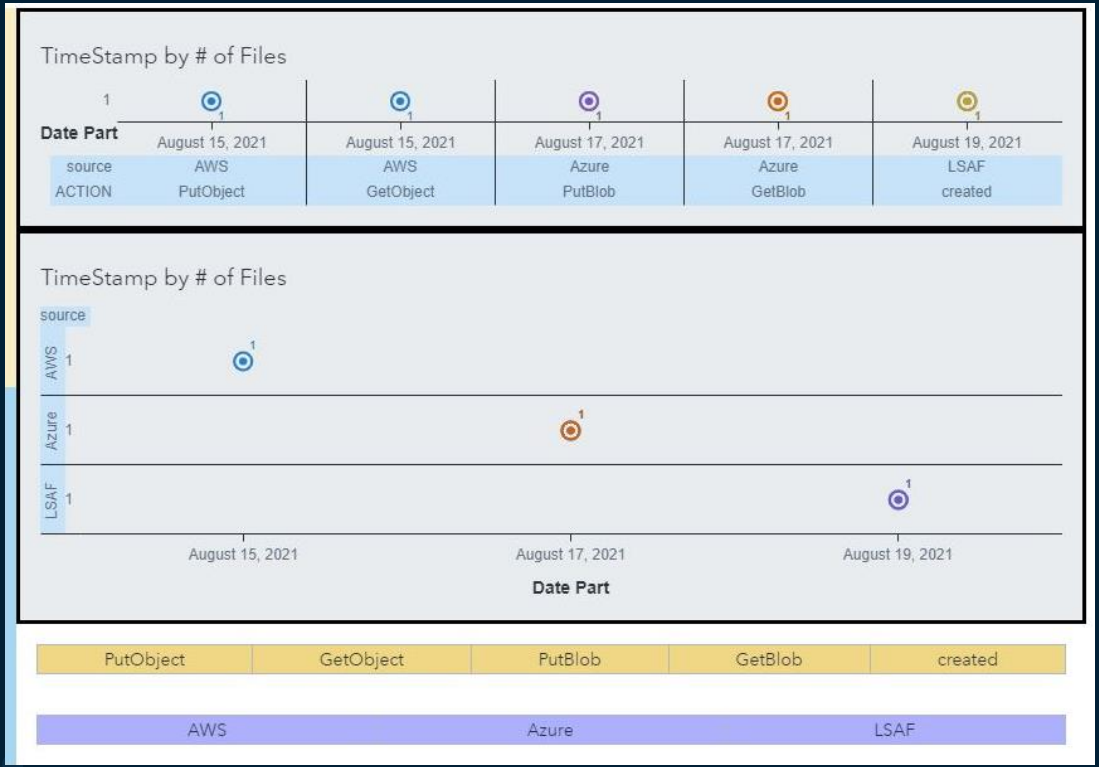
- Combining Audit Data across Cloud Environment provides data traceability

userAgent	source	action	userid	sourceLocation	sourceName
[rclone/v1.56.0]	AWS	GetObject	AROAY7OMEHNKKXPZDROA3:Sandeep.Juneja@SAS.com (AssumedRole)	hlsdata.s3.us-east-1.amazonaws.com	athena2.csv
[rclone/v1.56.0]	AWS	GetObject	AROAY7OMEHNKKXPZDROA3:Sandeep.Juneja@SAS.com (AssumedRole)	hlsdata.s3.us-east-1.amazonaws.com	athena2.csv
[rclone/v1.56.0]	AWS	GetObject	AROAY7OMEHNKKXPZDROA3:Sandeep.Juneja@SAS.com (AssumedRole)	hlsdata.s3.us-east-1.amazonaws.com	athena2.csv
[rclone/v1.56.0]	AWS	GetObject	AROAY7OMEHNKKXPZDROA3:Sandeep.Juneja@SAS.com (AssumedRole)	hlsdata.s3.us-east-1.amazonaws.com	athena2.csv
AzCopy/10.11.0 Azure-...	Azure	GetBlob	Sandeep.Juneja@SAS.com	https://lsafadls2.blob.core.window	000000.log
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	000000.log
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	000000.log
Mozilla/5.0 (Windows ...	Azure	PutBlob	Sandeep.Juneja@SAS.com	https://lsafadls2.blob.core.window	test_file.txt
Mozilla/5.0 (Windows ...	Azure	PutBlob	Sandeep.Juneja@SAS.com	https://lsafadls2.blob.core.window	readme.txt
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	readme.txt
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	readme.txt
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	readme.txt
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	readme.txt
rclone/v1.56.0	Azure	GetBlob	342fafc7-ac06-448d-982b-	https://lsafadls2.blob.core.window	readme.txt
	LSAF	created	sajune (Sandeep.Juneja@SAS.com)	/Users/sajune	athena.csv
	LSAF	created	sajune (Sandeep.Juneja@SAS.com)	/Users/sajune	athena2.csv
	LSAF	created	sajune (Sandeep.Juneja@SAS.com)	/Users/sajune	readme.txt

Ensuring Distributed Data custody on Cloud Platforms

Data Traceability across cloud Environments

- Combining Audit Data across Cloud Environment provides data traceability



Questions
Contact:

Sandeep.Juneja@biogen.com
Ben.Bocchicchio@sas.com

sas.com