Paper 4295-2020

Data Breaches Are Deadly – Let's Move Our Data with Cloud Data Exchange

Ivor G. Moan, SAS Institute Inc.

ABSTRACT

Data preparation involves collecting, processing, and cleansing data for use in analytics and business intelligence. These tasks include accessing, loading, structuring, purging, unifying (joining), adjusting data types, verifying that only valid values are present, and checking for duplicates and uniform data (for example, two birthdates for one person). When we look specifically at accessing data, we bump into the first real issue: dealing with sensitive and personal data. Many industries (for example, banks) have very strict rules concerning the movement, handling, and storage of data . Data preparation is playing the role of self-service data management. Traditional data management processes can produce data up to a point, but dynamic fine-tuning and last-minute work is being done in a self-service way, using data preparation tools. But how can we feed this process with sensitive data or personal data without potentially compromising this data?

Cloud Data Exchange is an intrinsic part of SAS® Data Preparation. With Cloud Data Exchange , we can give users secure, easy access to data, help users to easily manage access to remote data, to securely transport the data to any location (including the Cloud), enable users to control read and write access, and facilitate users to control and monitor data usage. Let's look at how Cloud Data Exchange can enable secure movement of data from source to on-premises, public cloud, private cloud, and hybrid cloud locations.

INTRODUCTION

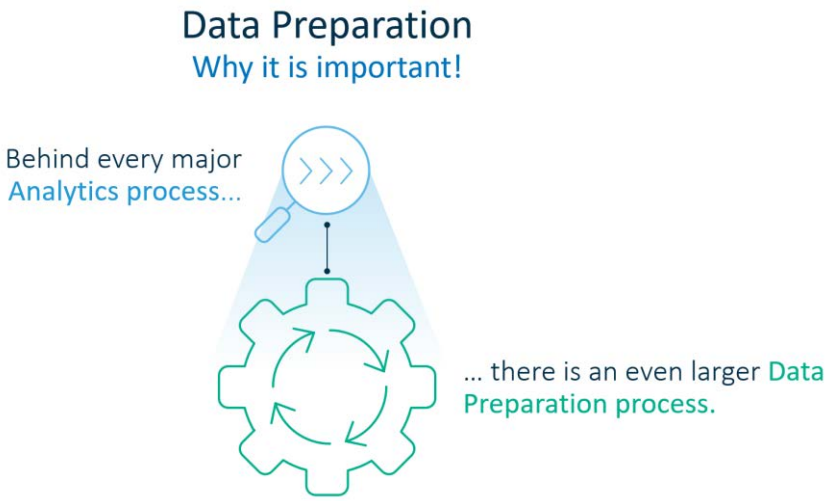

Figure 1. What is SAS Data Preparation and Why Does it Matter?

As the amount of data, and the number of sources increase, good data preparation is becoming both costlier and more complex. It is therefore the new paradigm that is shaping the market. It has also become, effectively, self-service data management. What is clear is that it is becoming more important to shape the data and to get it right for analytics. Many more companies are now data-driven. Businesses make decisions today based on data and it is vital to be able to access data quickly and prepare it for analysis. Big data environments such as Hadoop mean that it is impossible to move data from these environments. Instead, it is important to process data in place and combine resultant data with other sources as part of preparing data for analytics.

Data preparation is therefore an essential part of any analytics project. Getting the right data, and preparing it right, means that it is possible to get good answers to analytical questions. Use poor quality data, or data that have been badly prepared, and the results of your analysis are unlikely to be reliable.

But what if we want to feed our analytics project with sensitive data or personal data? How can we do this without potentially compromising this data?

HORROR STORIES

"Data breaches are deadly", what exactly does that mean? Here are some examples:

Figure 2. Horror of Horrors!

Equifax: In 2017, the credit reporting firm revealed that a breach from mid-May to the end of July 2017 potentially impacted 143 million US consumers. In March, it said that another 2.4 million people were possibly affected.

Anthem: The insurance giant saw tens of millions of client accounts compromised, with birthdates and other personal information released onto the black market.

Gundremmingen: This nuclear plant north of Munich discovered that their systems were infected with malware that could have given outside forces access to a system used for moving highly radioactive nuclear fuel rods.

SECURING YOUR DATA

There are three types of cloud solutions. Each of these offers a unique combination of advantages and drawbacks:

Public Cloud: These services offer accessibility and security. This security is most suitable for unstructured data, like files in folders. Most users do not receive a great deal of customized attention from public cloud providers. This option is affordable.

Private Cloud: Private cloud hosting services are on-premises solutions. Users assert unlimited control over the system. Private cloud storage is more expensive. This is because the owner manages and maintains the physical hardware.

Hybrid Cloud: Many companies choose to keep high-volume files on the public cloud and sensitive data on a private cloud. This hybrid approach strikes a balance between affordability and customization.

HOW SECURE IS CLOUD STORAGE?

All files stored on secure cloud servers benefit from an enhanced level of security. The security credential most users are familiar with is the password. Cloud storage security vendors secure data using other means as well:

- advanced firewalls
- intrusion detection
- event logging
- internal firewalls
- encryption
- physical security

The value is accessing the data and putting the data to use. But how do we do this without risking a data breach? What are the options when the data modeling and analysis is happening in the Cloud while the most relevant data is secured on in-premises systems? How can we securely, and equally important, easily access this data while preventing its exposure to the world?

Cloud Data Exchange is an intrinsic part of SAS Data Preparation. With Cloud Data Exchange, we can give users secure, easy access to data, help users to easily manage access to remote data, to securely transport the data to any location (including the Cloud), enable users to control read and write access, and facilitate users to control and monitor data usage.

Let's look at how Cloud Data Exchange can enable secure movement of data from source to on-premises, public cloud, private cloud, and hybrid cloud locations.

SOME SCENARIOS

A component part of Cloud Data Exchange is SAS® Data Agent. SAS Data Agent moves data for use by SAS Data Preparation and is optimized to connect to SAS Data Preparation that is running on a private or public cloud.

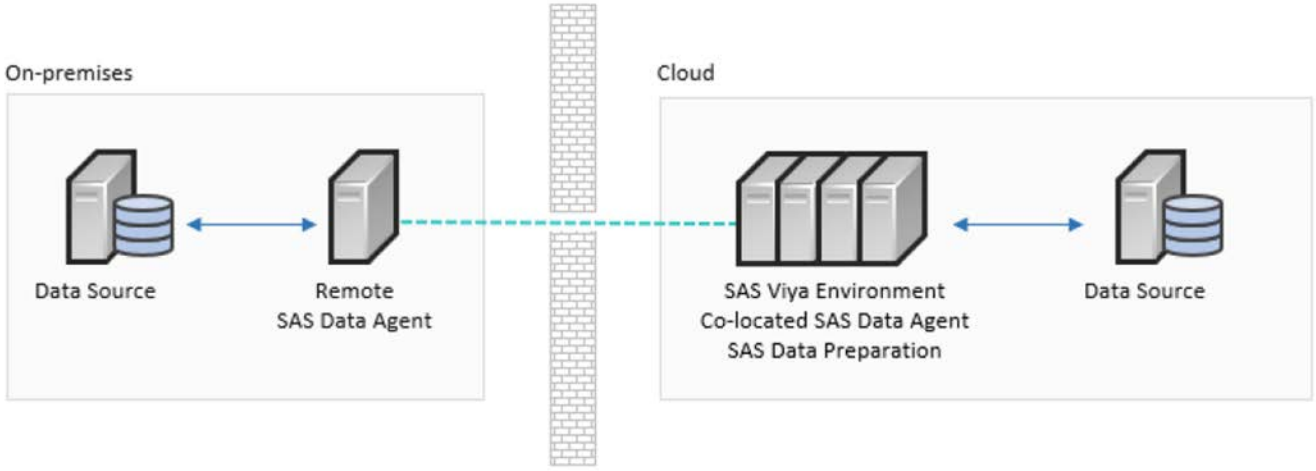

Figure 3. Remote and Co-Located SAS Data Agent Servers

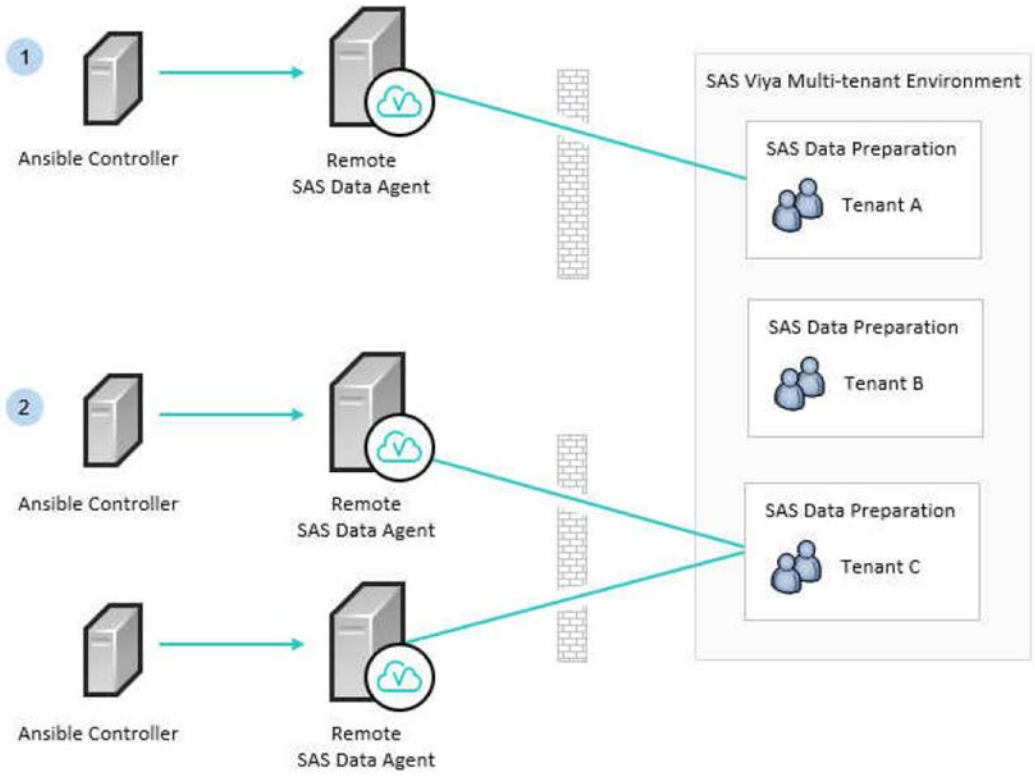

Figure 4. SAS Data Agent Servers Registered to Different Tenants

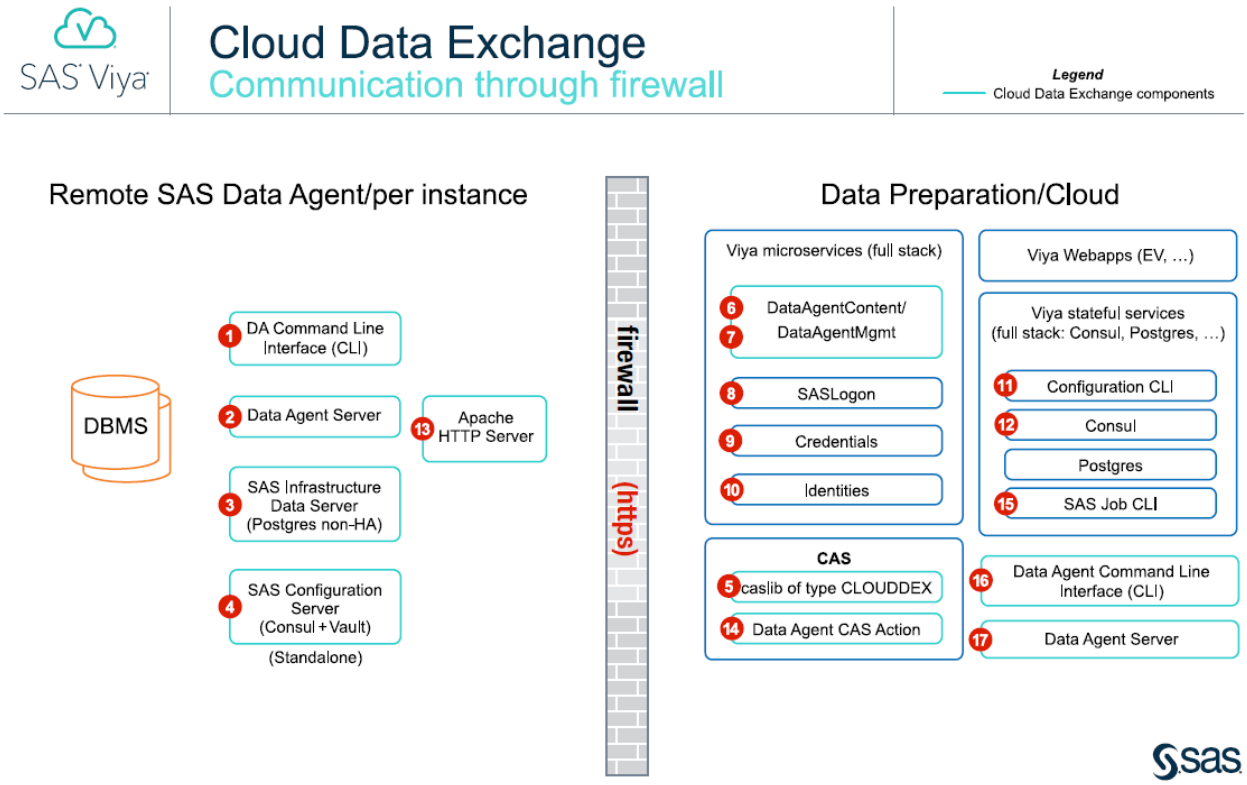

Figure 5. Cloud Data Exchange Communication Summary

1. SAS Data Agent Command Line Interface (CLI)
2. Remote SAS Data Agent Server
3. SAS® Infrastructure Data Server (PostgreSQL)
4. SAS® Configuration Server (Consul plus Vault)
5. caslib SAS® Data Connector CLOUDDEX
6. SAS® Viya® microservices: Data Agent Service (DataAgentContent)
7. SAS Viya microservices: Data Agent Service (DataAgentManagement)
8. SAS Viya microservices: SASLogon
9. SAS Viya microservices: Credentials
10. SAS Viya microservices: Identities
11. SAS Viya stateful services: Configuration CLI
12. SAS Viya stateful services: Consul
13. Apache HTTP Server
14. SAS Data Agent CAS Action
15. SAS® CLI Job
16. SAS Data Agent CLI
17. SAS Data Agent in the Cloud

Cloud Data Exchange is a data connection capability to securely access to on-premises data (behind a firewall) from a public or private cloud application. There can be on-premises data in various sources (like Oracle, Teradata, Hadoop, and so on). Cloud Data Exchange enables users to securely access these data sources from the Cloud, negotiates the on-premises firewall securely and responsibly, while transferring high volume data between on-premises sources and the cloud application. Cloud Data Exchange stores on-premises data source credentials (userid/password) in a secured vault. So, they never have to be stored or accessed outside the on-premises firewall.

Cloud Data Exchange allows access to databases that are not co-located with a SAS Viya deployment. This is accomplished by using secure standards-based communication that is coupled with sophisticated authentication and authorization models.

Cloud Data Exchange allows for the SAS Data Agent to reside outside of the rest of the SAS Viya deployment. The above graphic illustrates the components that support Cloud Data Exchange. This graphic also illustrates the split nature of the deployment. In this case, the description of SAS Data Agent refers to the portion of Cloud Data Exchange that resides inside a secured network. SAS Data Preparation in this topology indicates where the remainder of the SAS Viya deployment resides, which could be in a private or public cloud, or directly installed physically on-premises but in a different secured network domain.

SAS DATA PREPARATION

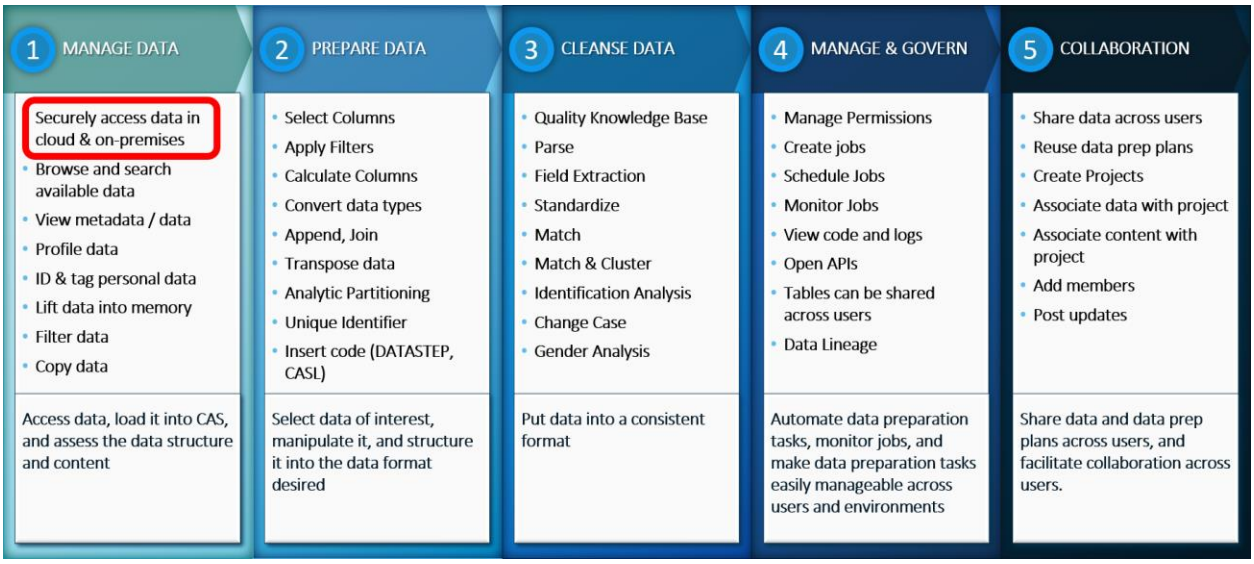

Figure 6. Cloud Data Exchange in the SAS Data Preparation Context

MOVING DATA FROM ON-PREMISES TO CLOUD

Challenges

- Insecure networks & threats
- Hackers want to steal data
- Exposed, Open ports
- Exposed UserIDs and Passwords
- Unencrypted data
- Unencrypted passwords
- Slow networks
- Long data transport times
- Diverse set of data sources
- Different database clients / APIs
- Difficulty to administer access to data
- Difficulty administering access to navigate the various Firewalls

What's needed

- Secure, reliable self service access to on-premises data
- Using standards for data encryption including TLS, and HTTPS
- Without opening ports
- Without giving users UserIDs and Passwords
- Encrypt data
- Encrypt all communications
- Parallelize data movement
- Centralize management of data sources
- Virtualize on-premises data to cloud users

TYPICAL SCENARIOS AND USE CASES

1. Customers who want to perform historical analysis on 10 years of sales data stored in their on-premises Oracle database
2. Customers who have been reluctant to convey data across their firewall for concerns over security
3. Customers who are in the cloud or are transitioning to the cloud
4. Customers with SAS 9x wanting to transition to cloud-based SAS Viya

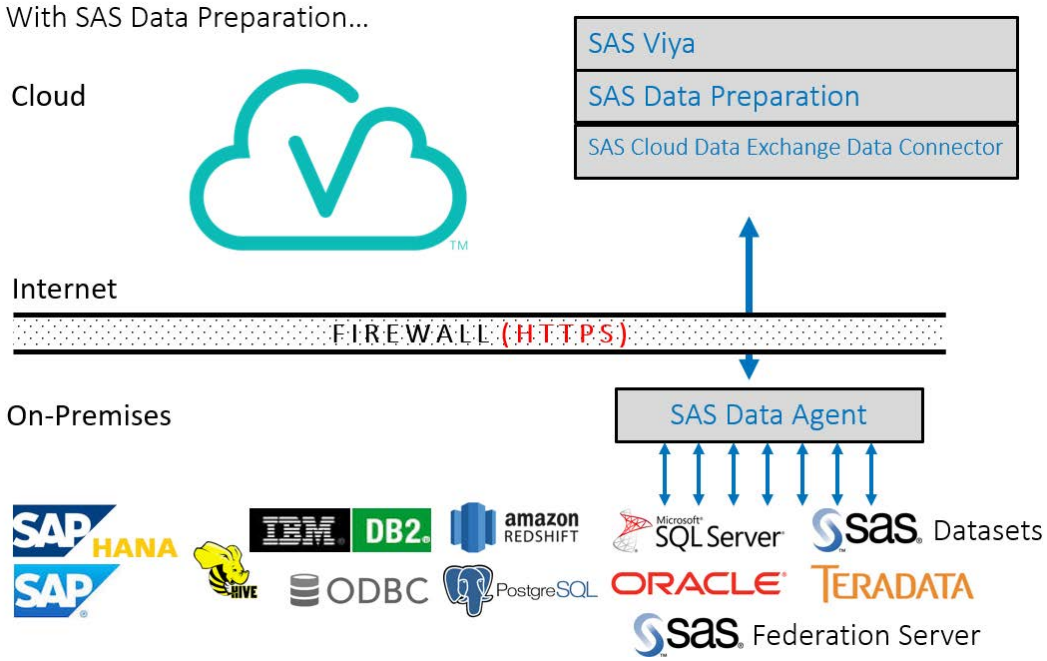

Figure 7. Cloud Data Exchange moving data to the cloud securely

SECURED COMMUNICATIONS

Cloud Data Exchange requires one open secured port rather than many in the firewall, thereby reducing typical data security concerns. An Apache HTTP Server is provided for this purpose. (See Figure 5, item 13.) With Cloud Data Exchange, SAS Data Agent can be placed in an organization's perimeter network, whereas the data resides on the LAN behind a firewall. This allows for monitoring of network traffic and isolation and protection of IT resources.

Any communication that occurs between the SAS Data Agent and SAS Data Preparation components, and any communication that occurs among the SAS Data Agent components is performed using industry-standard TLS encryption. This means that any sensitive data (including credentials) that must be moved between components is protected.

See *Encryption in SAS Viya: Data in Motion* for additional information (in support.sas.com).

CONFIGURATION – SAS VIYA

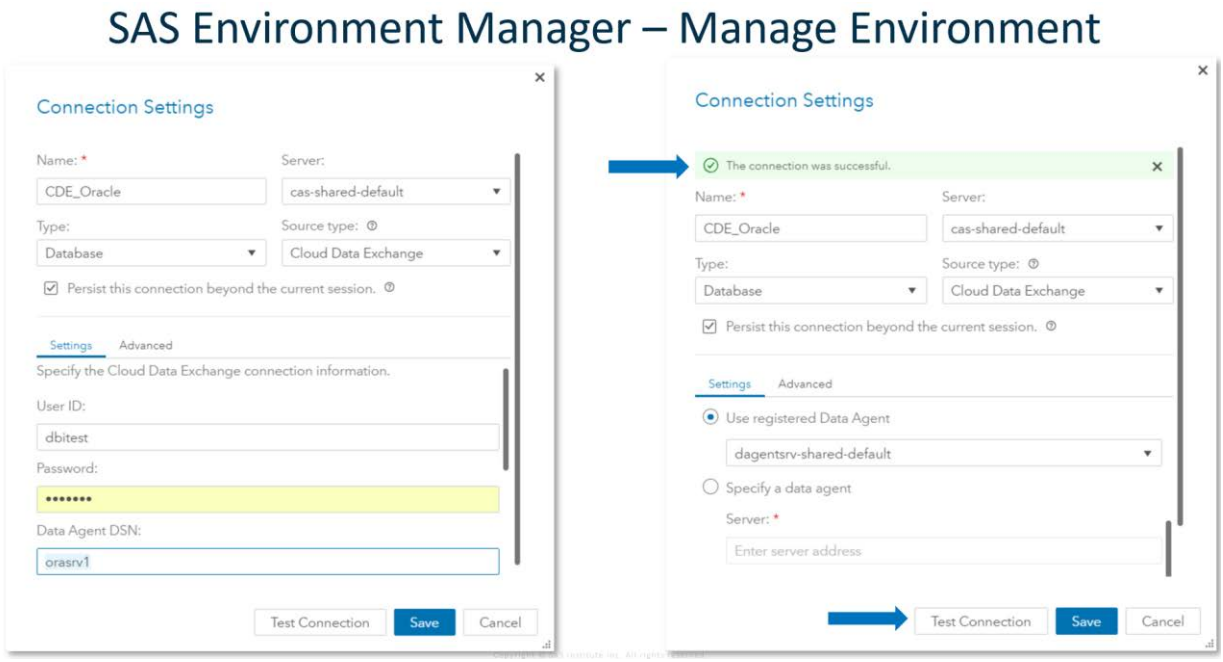

Figure 8. Defining a Global CASLIB in SAS® Environment Manager

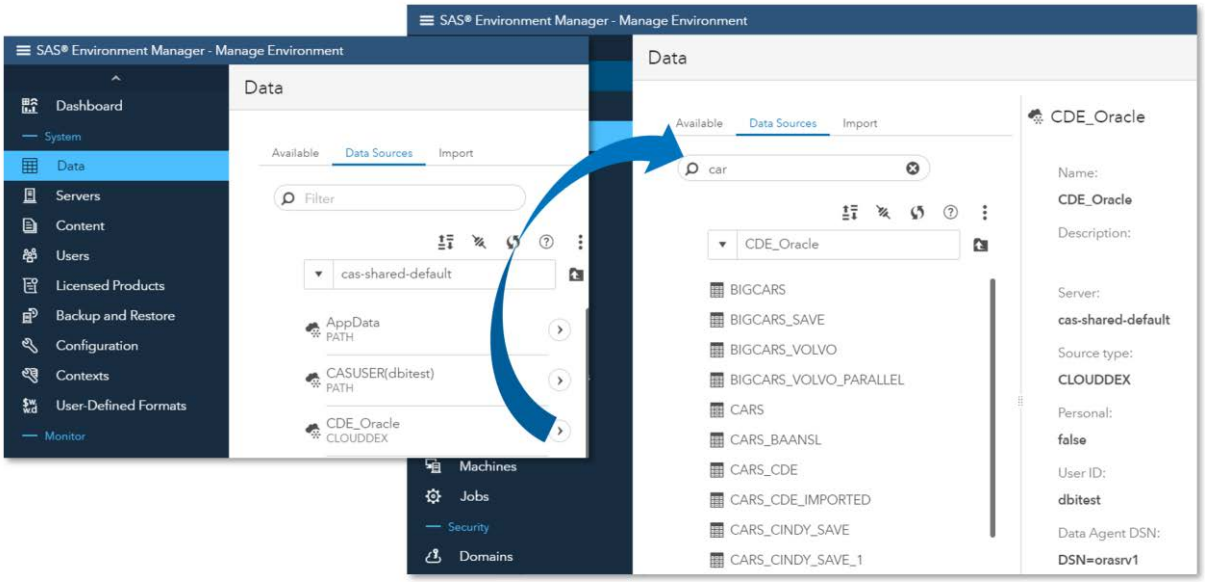

Figure 9. Finding Data in the Defined Data Source

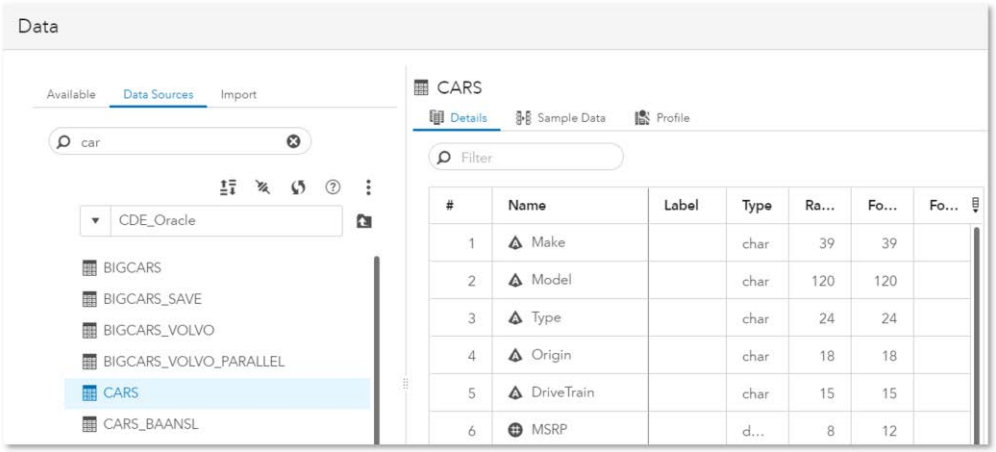

Figure 10. Data via the Cloud Data Exchange

RUNNING SOME CODE

```
1     OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
75
76    cas mySession sessopts=(caslib=casuser timeout=1800 locale="en_US");
NOTE: The session MYSESSION connected successfully to Cloud Analytic Services
      cloud-data-exchange-demo-cloud.dmmdev.sashq-d.openstack.sas.com using port 5570. The UUID is
      9860596c-ddd8-9b44-b4f9-835cc3c7ac2a. The user is eurigm and the active caslib is CASUSER(eurigm).
NOTE: The SAS option SESSREF was updated with the value MYSESSION.
NOTE: The SAS macro _SESSREF_ was updated with the value MYSESSION.
NOTE: The session is using 0 workers.
NOTE: 'CASUSER(eurigm)' is now the active caslib.
NOTE: The CAS statement request to update one or more session options for session MYSESSION completed.
77
78    proc cas;
79    session mySession;
80    action loadDatasource / name="clouddex"; run;
NOTE: Active Session now mySession.
NOTE: Cloud Analytic Services added the datasource 'clouddex'.
{clouddex}
81    quit;
NOTE: PROCEDURE CAS used (Total process time):
      real time           0.02 seconds
      cpu time            0.02 seconds
```

Figure 11. Activating the Cloud Data Exchange Data Connector in SAS Viya (CAS)

```
83    proc cas;
84       session mySession;
85       action addCaslib / caslib="dalib_ora"
86       session=false
87       datasource={srcType="clouddex",
88                   username="eurigm",
89                   password=XXXXXXXXXX,
90                   dataAgentName="dagentsrv-shared-default",
91                   catalog="ORCL_DSN",
92                   schema="TKTSTST1",
93                   conopts="dsn=orcl_dsn"
94                   };
95    run;
NOTE: Active Session now mySession.
NOTE: 'dalib_ora' is now the active caslib.
NOTE: Cloud Analytic Services added the caslib 'dalib_ora'.
96    quit;
NOTE: The PROCEDURE CAS printed page 1.
NOTE: PROCEDURE CAS used (Total process time):
      real time           0.06 seconds
      cpu time            0.05 seconds
```

Figure 12. Create a CASLIB to the Cloud Data Exchange Data Source Name (DSN)

```
76     proc cas;
77        session mySession;
78        action fileInfo / casLib="CDE_OraDSN"; run;
NOTE: Active Session now mySession.
NOTE: Connecting using logon 'eurigm'.
WARNING: Using server default session timeout of 3600
```

Figure 13. Getting File Information From a Cloud Data Exchange CASLIB

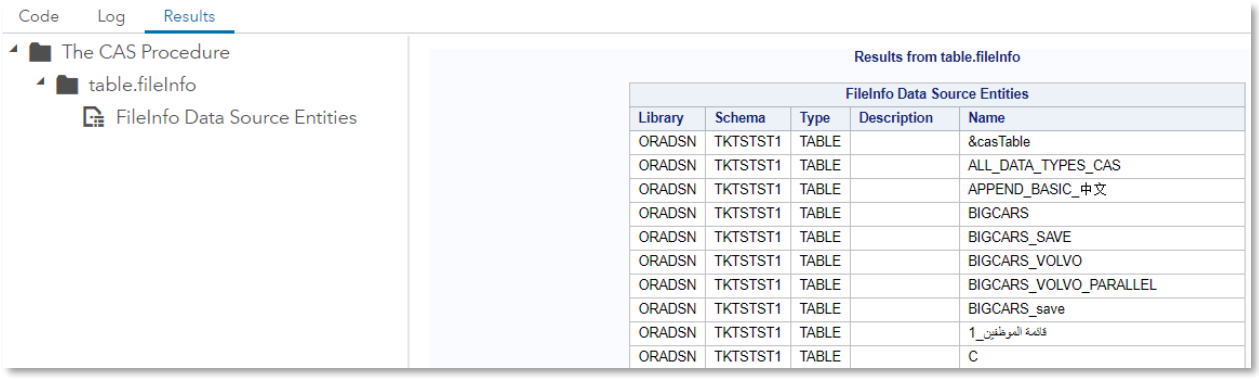

Figure 14. The Files and Tables That are Present

THE CHALLENGES THAT CLOUD DATA EXCHANGE ARE ADDRESSING

DATA TRANSPORT

How can data be securely copied to the Cloud?

> Any communication that occurs with Cloud Data Exchange components is performed using industry-standard TLS encryption.

How can only the necessary data be copied to the Cloud?

> Another vital capability of Cloud Data Exchange is its ability to process the data prior to securely transmitting result data.

PERSONAL DATA

Personal data is stored securely in the on-premises environment

How can this data be copied to the Cloud without exposing the personal data?

> Any communication that occurs with Cloud Data Exchange components is performed using industry-standard TLS encryption. Of course, it will now be transmitting result data securely to the Cloud, or to some other environment. There will be a requirement to secure the target location equally well. However, Cloud Data Exchange allows the processing of SAS FedSQL views prior to the transmission of the result data. Within these views we can pseudonymize, anonymize, and/or encrypt the personal data, thus protecting it at the source.

PERFORMANCE

How can the appropriate data be extracted from the huge quantities of on-premises data?

> Cloud Data Exchange allows the processing of SAS FedSQL views prior to the transmission of the result data. These views, indeed, can result in implicit/explicit pass-through of query code to the underlying source systems. Data can be joined and subset prior to transmission resulting in massive performance gains and efficiencies.

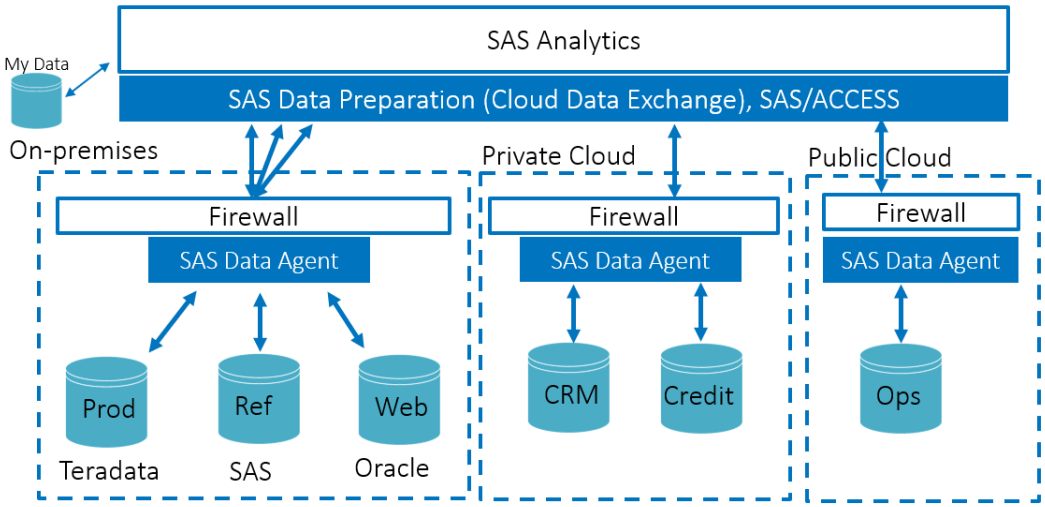

Figure 15. Cloud Data Exchange Connectedness

SECURING ACCESS IN SAS VIYA

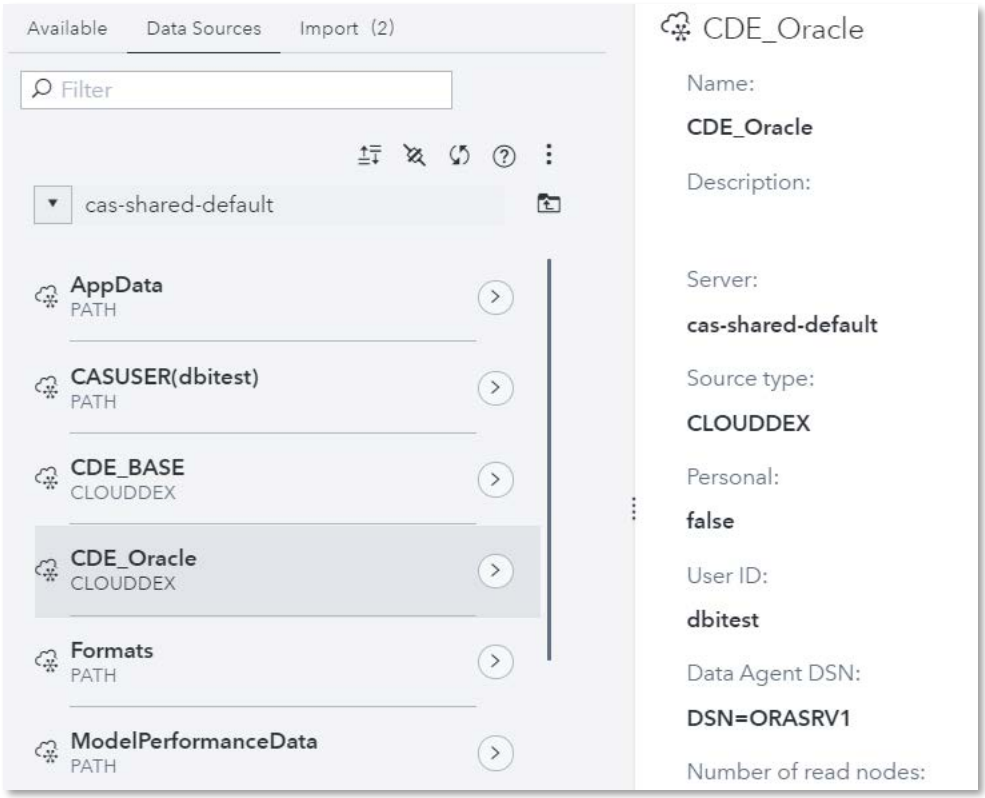

Figure 16. We Have a Cloud Data Exchange Data Source (CASLIB) On-Premises

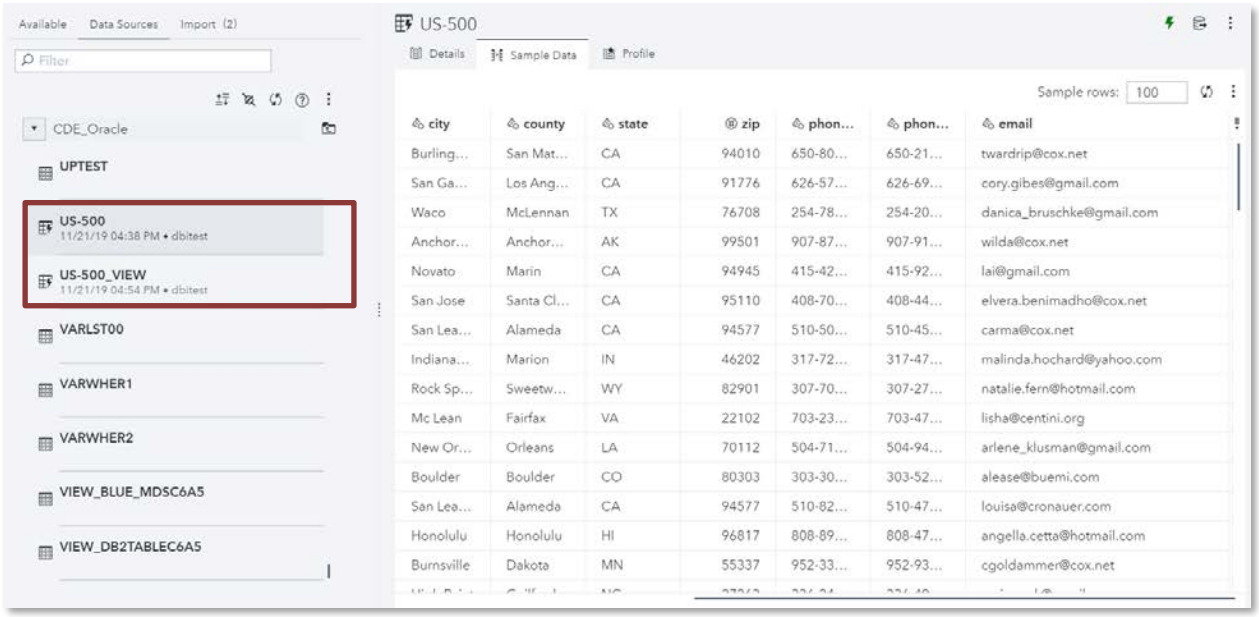

Figure 17. In the CASLIB, We Have a Table and a View of This Table Must Be Secured

Figure 18. Regular Users Have No Access to the Table – An Administrator Account Has Full Control

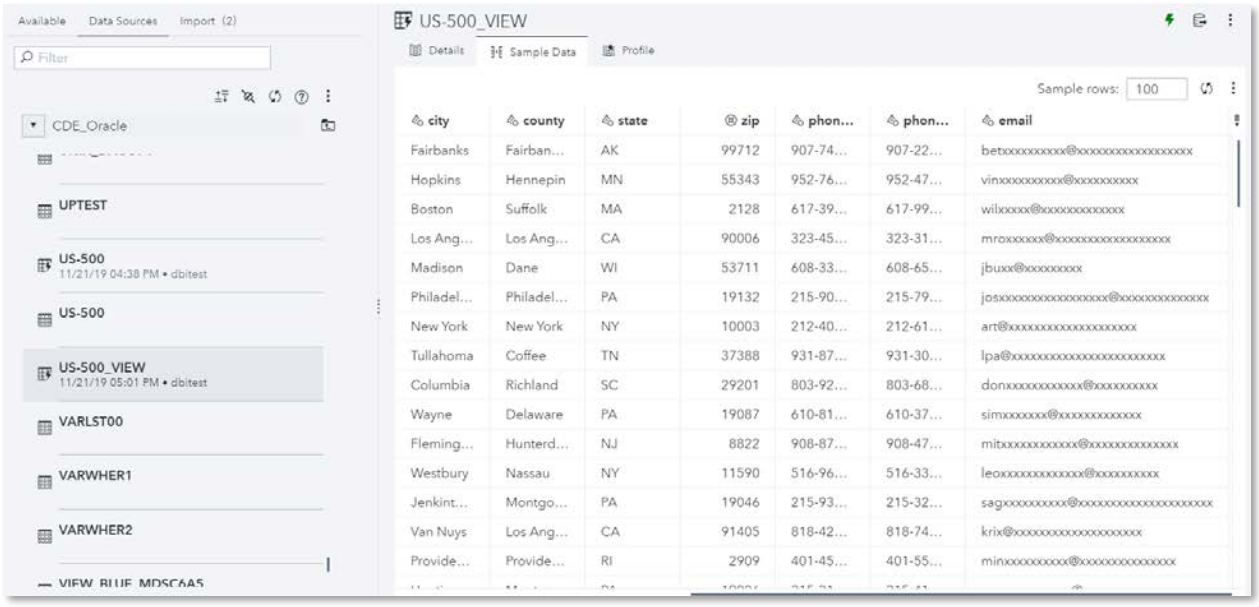

Figure 19. SAS FedSQL View of the Table is Masking the Email Address

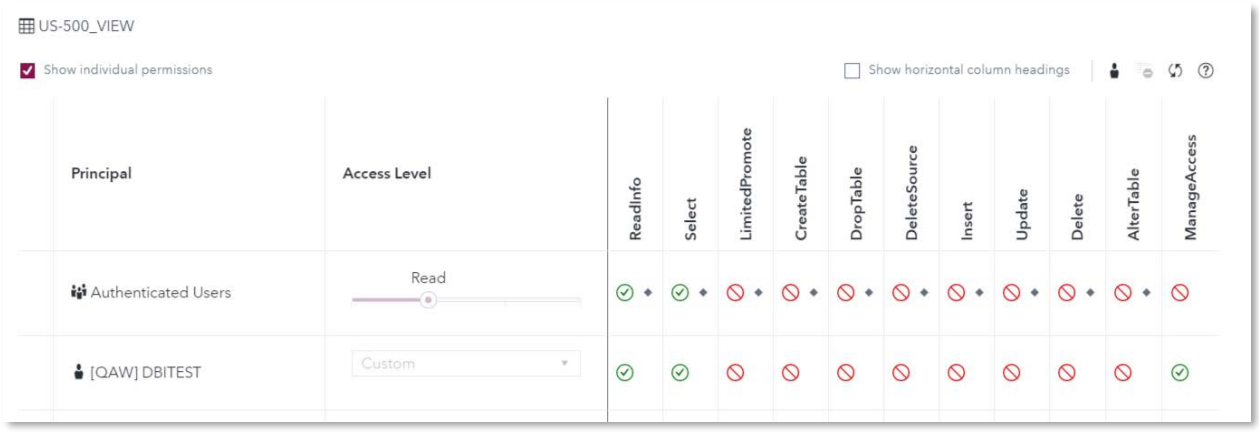

Figure 20. Regular Users Have Been Granted Some Access to the SAS FedSQL View

CONCLUSION

Dealing with sensitive and personal data requires secure data handling and transmission. SAS Data Preparation is deeply embedded in SAS Viya – the capability is called Cloud Data Exchange. Many industries have very strict rules concerning the movement, handling, and storage of data (for example, banks). Cloud Data Exchange ensures that all communication (users, passwords, and data) is securely handled using industry-standard TLS encryption.

Cloud Data Exchange is much more than a means to transmit data securely. It is a mechanism that allows users access to secure data platforms without passing around user ids and passwords. It is a mechanism that allows users to transparently access data from secure platforms just like any other data source. It is a mechanism that allows for the orchestration and determination of where processing and data filtering will occur, facilitating efficiency and optimal performance. It is a mechanism that facilitates the transition from on-premises operations to cloud or hybrid environments.

The creation and deployment of SAS FedSQL views facilitates the processing of data behind the firewall on either the SAS Data Agent, or on the source system itself. This is particularly useful when dealing with huge amounts of data. In this scenario, processing, filtering, and masking of sensitive data becomes a compelling capability that transparently operates in the background. After all, users just want to get the data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ivor G. Moan
SAS Institute Inc.
In der Neckarhelle 162
69118 Heidelberg, Germany
Email: Ivor.Moan@sas.com
Web: www.sas.com