Implementing SAS®9.4 Software in Cloud Infrastructures

Cloud Enablement Virtual Team
Release Information

Content Version: 1.0 July 22, 2016.

Trademarks and Patents


SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.
# Contents

1. Introduction ........................................................................................................................................... 1  
   1.1. Purpose of This Technical Paper ................................................................................................. 1  
   1.2. What This Technical Paper Covers .............................................................................................. 1  
   1.3. What This Technical Paper Does Not Cover ............................................................................... 1  
2. Summary.................................................................................................................................................. 1  
3. Introduction to Cloud Computing ........................................................................................................ 2  
   3.1. Cloud Providers ............................................................................................................................ 3  
   3.2. Differences between SAS®9.4 Software on Premises and in the Cloud .................................. 3  
   3.3. Benefits of SAS®9.4 Software Cloud Deployments .................................................................... 3  
   3.4. Risks Associated with SAS®9.4 Software Cloud Deployments ................................................. 4  
4. Expectation Settings ............................................................................................................................... 4  
   4.1. Common Misconceptions .............................................................................................................. 4  
   4.2. Statement of Support ..................................................................................................................... 4  
   4.3. IaaS or SaaS .................................................................................................................................. 5  
   4.4. Integration with Cloud Services .................................................................................................... 5  
   4.5. Respective Responsibilities .......................................................................................................... 5  
5. Implications for SAS®9.4 Software Licensing ................................................................................... 7  
   5.1. Licensing of SAS®9.4 Software in AWS ..................................................................................... 7  
   5.2. Licensing for SAS®9.4 Software with SAS® Grid Manager ....................................................... 7  
6. Costs ....................................................................................................................................................... 7  
   6.1. Infrastructure Costs ....................................................................................................................... 7  
   6.2. Cloud Server Costs ....................................................................................................................... 8  
   6.3. Cloud Storage Costs ..................................................................................................................... 8  
   6.4. Data Transfer Costs ...................................................................................................................... 8
7. General Considerations for SAS®9.4 Software Cloud Deployments

7.1. Selecting a Cloud Provider
7.2. Connectivity
7.3. Impact of SAS Client Locations

8. Scalability in the Cloud

8.1. Scale Up
8.2. Scale Down
8.3. Scale Out
8.4. Scale In
8.5. Elasticity
8.6. Implications of Scaling SAS®9.4 Software

9. Architectural Best Practices

9.1. Authentication
9.2. Create VPC
9.3. Server Locations
9.4. HA
9.5. D/R
9.6. Maximizing Network Throughput within the Cloud with AWS Placement Groups
9.7. Storage Selection
9.8. Grid-Specific Information
9.9. Monitoring and Auditing Cloud Deployments

10. Backing Up and Restoring SAS®9.4 Software in Cloud Infrastructures

10.1. What to Back Up
10.2. Backing Up
10.3. Restoring
10.4. Automation
List of Tables and Figures

Table 1. Example of SAS®9.4 Software Limitation ................................................................. 4
Table 2. SAS®9.4 Software Integration ................................................................................... 5
Figure 1. High-Level Plan for SAS®9.4 Software Cloud Deployments ...................................... 6
Table 3. Pricing Policy ............................................................................................................... 9
Figure 2. On-Premises Database Scenario ................................................................................ 10
Figure 3. Target Database in Same Cloud as SAS®9.4 Software Scenario ................................. 11
Table 4. Selecting Cloud Provider Criteria ............................................................................. 11
Figure 4. User, Corporate Server, and Cloud Zones ................................................................. 13
Figure 5. Site-to-Site Connection ............................................................................................. 14
Figure 6. Point-to-Site Connection ......................................................................................... 15
Table 5. Private Subnet Settings .............................................................................................. 16
Table 6. Public Subnet Settings ............................................................................................... 16
Table 7. SAS Client Locations ................................................................................................. 17
Figure 7. Scaling Up ................................................................................................................ 18
Figure 8. Scaling Down .......................................................................................................... 18
Figure 9. Scaling Out .............................................................................................................. 19
Figure 10. Scaling In .............................................................................................................. 19
Table 8. Icon Meanings ........................................................................................................... 20
Table 9. SAS® Metadata Server Deployment ........................................................................ 21
Table 10. SAS Web Server Deployment ................................................................................ 21
Table 11. SAS Web Application Server Deployment ............................................................... 22
Table 12. SAS Compute Server Deployment ......................................................................... 23
Table 13. SAS® Visual Analytics LASR™ Nodes Deployment .................................................. 23
Table 14. SAS®9.4 Software Deployment Server Locations across Availability Zones/Availability Sets ........................................................................................................... 26
Table 15. Storage Types ........................................................................................................ 27
1. Introduction

1.1. Purpose of This Technical Paper

This technical paper describes cloud concepts and relates these concepts to SAS® 9.4 software deployments, and it should be read in preparation for SAS software cloud deployments using the Infrastructure as a Service (IaaS) cloud model. The pros and cons of SAS software cloud deployments are covered by this paper, and it also provides information regarding SAS recommendations. Lastly, this paper identifies pitfalls to be avoided when conducting SAS software cloud deployments.

1.2. What This Technical Paper Covers

The topics covered in this technical paper relate specifically to SAS software deployments that utilize SAS® Metadata Servers, SAS middle-tier servers, and SAS compute servers, such as those involving SAS® Visual Analytics, SAS® Visual Statistics, SAS® Office Analytics, and SAS® Data Management. While there are many cloud providers in existence, the concepts and recommendations in this paper cover the most common cloud providers such as Amazon Web Services (AWS) and Microsoft® Azure, as well as on-premises cloud infrastructures such as OpenStack.

1.3. What This Technical Paper Does Not Cover

The content of this technical paper does not relate to the SAS®9.3 platform or earlier releases of SAS software, SAS® Cloud, Platform as a Service (PaaS), or Software as a Service (SaaS).

2. Summary

SAS®9.4 software can be deployed in the cloud. While many considerations and design decisions are the same as those for SAS software on-premises deployments, there are also some differences. The following points summarize the key discussions in this technical paper:

- The overall process for SAS software cloud deployments is essentially the same as it is for SAS software on-premises deployments:
  - SAS infrastructure sizings and architectures must be created.
  - SAS software licensing metrics are the same for SAS software cloud deployments as compared with SAS software on-premises deployments.
  - Cloud administration, security, and monitoring are the responsibility of clients.
  - While SAS software can be scaled, care must be taken for clients to remain within licensing agreements.
  - In most cases, scaling a SAS infrastructure results in an outage of service during the process.
  - Topics such as high availability (HA), disaster recovery (D/R), and back-up and restoration are as important in SAS software cloud deployments as they are in SAS software on-premises deployments.

- Choosing a cloud provider is critical:
  - Not all cloud providers offer the same set of services.
  - For a given cloud provider, not all locations provide all services, so it should be verified that the desired services are available in the preferred location.
Local laws and regulations can affect what may put into the cloud (e.g., certain geographies may restrict the storage and processing of data in a cloud location out of country or state).

- The cost of a cloud infrastructure is a core consideration.

- Security, networking, and data location considerations should be taken into account:
  - Server instances should not be accessible over the public Internet, and third-party access should be managed via a VPN.
  - The connectivity between a cloud customer and a cloud provider is a key area to be considered and addressed.
  - Moving data out of the cloud incurs data transfer charges for cloud customers.

3. Introduction to Cloud Computing

Cloud computing is a relatively new concept and as such it continues to evolve and expand in terms of capabilities and offerings both in the commercial and open source realms. According to the National Institute of Standards and Technology (NIST), cloud computing is defined as follows:

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

In other words, cloud computing differs from traditional methods followed by organizations in deploying information technology assets and solutions. Whereas lengthy lead times for the design, procurement, and on-boarding of new infrastructures had previously been the norm, the cloud enables extremely fast access to new resources for use by organizations. There is also a paradigm shift away from having to purchase infrastructure under capital expenditure budgets in order to meet demand for the lifetime of an application or infrastructure, which was typically the order of magnitude of a few years. Cloud computing allows cloud customers to begin with initial small deployments and then scale when required so that the size of the infrastructure is always matched to the demands placed upon it. Because access to cloud infrastructures is “rented” from cloud providers to organizations, the costs associated with cloud infrastructures can be considered an operational expenditure. The ability to attribute IT infrastructure costs to operational expenditure provides real benefits from a financial strategy and management perspective for many organizations. NIST states that a cloud computing model consists of five essential characteristics, three service models, and four deployment models. Of the essential characteristics, the ones most pertinent to SAS®9.4 software cloud deployments are as follows:

- **Resource pooling** – organizations typically do not have control over where their cloud infrastructures run except for at a high level that specifies a provider’s data center; cloud providers allocate and re-allocate based on overall demand of the physical infrastructure used to support cloud customers’ services

- **Rapid elasticity** – infrastructure and application capabilities can be scaled rapidly in order to meet demand, often in an automated manner.

From a service model perspective, the two pertinent ones from a SAS software perspective are as follows:

- **IaaS** – SAS software cloud deployments are the same as SAS software on-premises deployments, the only difference being that the SAS software runs on a cloud infrastructure instead of on an infrastructure owned by the client

- **SaaS** – SAS® Cloud Analytics is a set of defined offerings for clients wishing to leverage a SaaS model for SAS software; not all SAS software is available via this model.
Of the deployment models, the *private cloud* is the focus of this technical paper. NIST defines private clouds as follows:

“Infrastructure (that) is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.”

The key attribute is that private clouds can be either on or off premises. Many organizations that have outsourced their IT infrastructures to external organizations may believe that they already have an on-premises private cloud based on this definition. However, unless an infrastructure meets all essential characteristics, it most likely is not a true private cloud. Most often, the characteristic that is not satisfied is rapid elasticity.

### 3.1. Cloud Providers

Clients have the choice of deploying SAS software in either an off-premises cloud or an on-premises cloud. The following is a list of cloud providers for each of the two options:

- **Off-premises cloud:**
  - AWS
  - Google Cloud Platform
  - IBM SoftLayer
  - Microsoft Azure
- **On-premises cloud:**
  - OpenStack
  - Microsoft Azure stack.

### 3.2. Differences between SAS®9.4 Software on Premises and in the Cloud

Over the history of computing, there have been five major macro trends in computing, as follows:

1. Monolithic computers with connected terminals
2. PC revolution moved processing to desktops
3. Consolidation to client–server model with mid-range servers
4. Consolidation of servers via virtualization (on premises)
5. Cloud computing is the next evolutionary step, moving servers (and some clients) to a service-based virtualization layer, most often off premises.

### 3.3. Benefits of SAS®9.4 Software Cloud Deployments

There are several benefits to SAS software cloud deployments. These can be summarized as follows:

- Perfect for short proofs of concept or projects because an infrastructure can be rented for the duration of a project and there are no lead times for hardware
- Ability to grow a SAS infrastructure without having to buy physical hardware
- Ability to save on infrastructure costs, especially if a SAS infrastructure does not have to be on all the time or by right-sizing a SAS infrastructure for only current demand
- Can be deployed with very short lead times compared to purchasing a physical infrastructure
- Ability to deliver a solution rapidly can have significant business benefits.
3.4. Risks Associated with SAS®9.4 Software Cloud Deployments

The following provides areas where clients must pay attention in order to ensure smooth SAS software cloud deployments:

• Require personnel to have cloud administration skills; creating a solid cloud infrastructure involves many areas of specialization:
  – Network design and management
  – Network security
  – Operations management to ensure that a cloud infrastructure runs smoothly with minimal to no unplanned downtime
• Ensure that variable costs associated with cloud deployments can be managed
• Be aware of any legal and statutory requirements regarding the storing and processing of data in a cloud infrastructure; the jurisdiction under which a client operates can materially impact its ability to leverage the cloud, and legal advice should be sought if there is any doubt
• Ensure that there is a secure connection to the cloud
• Ensure that a dedicated private VPN link to the cloud provider exists (mandatory).

4. Expectation Settings

4.1. Common Misconceptions

Common misconceptions about cloud deployments are as follows:

• IaaS cloud providers cannot help with the deployment and management of software in the cloud.
• If there is software that does not scale well, an IaaS cloud would not make it more scalable.
• If a software deployment has not been designed for HA, IaaS cloud would not make it HA.
• Deploying SAS®9.4 software in the cloud does not simplify the SAS software deployment process overall; the process of installing SAS software in the cloud is the same as installing it on premises.

4.2. Statement of Support

SAS software is supported for virtualization environments. Thus, if a client chooses to use an IaaS cloud provider such as AWS or OpenStack, that statement of support is extended to the cloud provider assuming that the provider fulfills the requirements of the specific SAS software. Not all SAS software is supported in cloud infrastructures or have additional requirements that clients must adhere to. Therefore, clients should consult SAS to determine whether or not their SAS software implementations are supported prior to attempting to deploy them in the cloud. One example of such a limitation is shown in Table 1.

<table>
<thead>
<tr>
<th>Solution</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAS® Grid Manager</td>
<td>![Warning] While SAS® Grid Manager is supported in cloud infrastructures, dependency on having efficient shared clustered file system available does exist</td>
</tr>
</tbody>
</table>

*Table 1. Example of SAS®9.4 Software Limitation*
4.3. IaaS or SaaS

This technical paper focuses on topics that relate to SAS software private cloud deployments using the IaaS deployment model.

4.4. Integration with Cloud Services

In addition to infrastructure, cloud providers also sell a wide range of applications via SaaS models. Table 2 lists a number of common examples and indicates how and if SAS software integrates with them.

<table>
<thead>
<tr>
<th>SaaS product</th>
<th>SAS®9.4 software integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS Redshift</td>
<td>SAS®9.4 software can access data via SAS/ACCESS® Interface to ODBC together with Redshift ODBC driver, or via SAS/ACCESS® Interface to Amazon Redshift</td>
</tr>
<tr>
<td>AWS Relational Database Service</td>
<td>SAS®9.4 software can access data via standard SAS/ACCESS® Interface engines together with native client (and Java™ Database Connectivity driver for SAS Web applications) for Oracle, PostgreSQL, SQL Server, and MySQL databases</td>
</tr>
<tr>
<td>Azure SQL Database</td>
<td>SAS®9.4 software can access data via SAS/ACCESS® Interface to Microsoft SQL or SAS/ACCESS® Interface to ODBC (and Java™ Database Connectivity driver for SAS Web applications)</td>
</tr>
<tr>
<td>AWS Elastic Map Reduce</td>
<td>SAS®9.4 software can access via SAS/ACCESS® Interface to ODBC; SAS/ACCESS® Interface to Hadoop is not currently supported</td>
</tr>
<tr>
<td>AWS S3 (object storage)</td>
<td>SAS®9.4 software can read data from object storage using HTTP file access method</td>
</tr>
<tr>
<td>AWS Simple Active Directory</td>
<td>Can be used as authentication provider for SAS®9.4 software; additional configuration is required to enable users to log in with usual Microsoft® Windows® credentials</td>
</tr>
<tr>
<td>Azure Active Directory and AWS Directory Services</td>
<td>SAS®9.4 software interfaces via LDAP or LDAPS protocols; neither Azure Active Directory nor AWS Directory Services provides LDAP interface and thus are not suitable as directory services for SAS®9.4 software</td>
</tr>
<tr>
<td>Azure HDInsight</td>
<td>Not supported for Hadoop distribution on SAS®9.4 platform</td>
</tr>
</tbody>
</table>

Table 2. SAS®9.4 Software Integration

SAS software can be integrated with cloud-hosted Active Directory services from AWS and Microsoft Azure. The recommended options that enable users to log in to a SAS infrastructure with their usual Microsoft Windows® credentials are as follows:

- Integrate directly with Active Directory on-premises deployments via VPN
- Implement a specific federated Active Directory architecture in the cloud.

4.5. Respective Responsibilities

At a high level, Figure 1 on the next page indicates the responsibilities of the various teams involved in an implementation project for SAS software cloud deployments. This process is identical to what would be expected for SAS software on-premises deployments. SAS software cloud deployments are therefore expected to require the same level of resourcing. The validation testing phase typically run by a client’s business and IT teams may require...
more effort than SAS software on-premises deployments, especially if this is one of the first cloud deployments for that particular client.

4.5.1. Client Responsibilities

Clients are responsible for standing up infrastructures onto which SAS software can be deployed. This entails creating some or all of the following:

- Virtual networks
- Sub-networks (subnets)
- Firewall rules
- VPN connections
- Servers and instances
- Storage attachments
- Integration of SAS infrastructure with authentication provider (e.g., configuration of PAM on Linux with Active Directory)
- SAS infrastructure pre-requisites:
  - Partitions
  - Users and groups
  - Folders and permissions
  - Operating system package updates
  - Installation of third-party software, such as required database clients
  - Tuning of operating system settings
4.5.2. SAS Responsibilities

We work with clients through a series of architecture and/or design workshops in order to create a customized SAS architecture design document that describes the target state architecture for a SAS software cloud deployment. In cases where a client has engaged us to deploy the SAS software, we would perform the following tasks:

- Install SAS server software
- Configure SAS server software
- Install SAS client software as agreed upon in the services statement of work
- Perform validation testing and document the results
- Hand the SAS infrastructure over to the client.

5. Implications for SAS®9.4 Software Licensing

For SAS®9.4 software cloud deployments, the licensing model is the same as for SAS software on-premises deployments. If the SAS software is priced based on server capacity, the fees would be priced by the total processor cores available to the selected cloud instance type.

5.1. Licensing of SAS®9.4 Software in AWS

For SAS software licensed according to server CPU capacity, the number listed in the "vCPU" column associated with the Amazon Instance type being licensed is divided by 2 in order to arrive at the total processor core count, which is the number required for licensing purposes (2-core license minimum). For example, consider an AWS EC2 m4.2xlarge instance type, which has 8 vCPUs (4-core license required). This is due to the way that AWS hyper-threads cores: an instance with 32 vCPUs is equivalent to a 16-core server. If a client elects to run an Amazon virtual private cloud (VPC) or is using the elastic IP feature with fixed host names, it can choose to provide us with the host name and/or IP address assigned to its Amazon cloud instance, thereby adding an extra layer of security to the license. SAS account executives should be consulted prior to SAS software cloud deployments.

5.2. Licensing for SAS®9.4 Software with SAS® Grid Manager

The licensing requirements for a SAS® Grid Manager cloud deployment are exactly the same as for a SAS software on-premises deployment.

6. Costs

6.1. Infrastructure Costs

SAS®9.4 software on-premises deployments have well-defined costs associated with them. The number and size of servers, the amount of storage, and the number of environments for current and known future growth over a defined period of time is locked in up front. Because the infrastructure is supplied by clients, hardware costs are also calculated up front. In addition, there are operational costs associated with the management of the SAS
infrastructure, SAS software and third-party licensing costs, and personnel costs associated with the operation and administration of the SAS infrastructure.

The move to cloud computing provides challenges to clients because infrastructure costs are based on usage instead of being static amounts. If the elasticity capabilities of the cloud are leveraged by a client for running SAS software, it follows that the number and size of servers, the amount of storage, and even the number of active environments could change on a regular and/or frequent basis. Every change to the SAS infrastructure affects the cost to clients based on the pay-as-you-go pricing model employed by cloud providers. This should be explored carefully and is the responsibility of a client when choosing a particular cloud provider. These considerations can also become part of the design of SAS software cloud deployments, where one could choose to prioritize performance over infrastructure costs, or vice versa.

6.2. Cloud Server Costs

In order to minimize the cost of cloud infrastructures, there are a number of strategies that can be implemented by cloud customers. These strategies are discussed in the following subsections.

6.2.1. Shut Down When Not In Use

If a PROD environment is used only during business hours, shutting the servers down overnight is a good way to reduce costs. This process should be automated so that the SAS infrastructure is ready to work when the client's users begin their workdays. Cost savings can similarly be realized for non-PROD environments. For example, TEST environments should be provisioned only when they are required and should be turned off when they are not in use. Once test cycles have been signed off and content has been promoted to a PROD environment, TEST environments could theoretically even be destroyed.

6.2.2. Using Reserved Instances

Many cloud providers offer large discounts for reserved instances. The hourly price can be up to half that of the on-demand cost. This can make sense for PROD environments, where licenses are annual. Further savings can also be achieved by shutting down environments when they are not in use.

6.3. Cloud Storage Costs

Cloud providers have a number of different storage offerings. Each type of storage has different characteristics regarding capacity limits, speed, and persistence, as well as differing cost policies. Shutting down a server does not delete its associated storage, and storage costs (unlike the costs associated with the instance itself) continue to accrue even when servers are off. Storage that is initialized upon re-start of a server (ephemeral) is clearly not a good option for application software or permanent data, but is a great choice for temporary storage such as SASWORK and UTILLOC. Some types of storage can cost more, based on the number of I/O operations per second (IOPS) being consumed. When designing a SAS infrastructure that has heavy I/O requirements, clients should select instance types that provide multiple ephemeral drives at low or zero cost. These ephemeral drives can be striped into high-performance RAID 10 or RAID 0 arrays, which is ideal for SASWORK.

6.4. Data Transfer Costs

None of the cloud providers listed in Table 3 charge for inbound data coming into their clouds, but each one has its
own pricing policy for moving data out of their clouds or between their cloud centers in various geographical locations (outbound). This can have a drastic effect on price when different pieces of software must exchange vast amounts of data. Such is the case with SAS software, in which a common use is to extract large amounts of data out of databases. Table 3 contains links to the pricing information for four common cloud providers.

<table>
<thead>
<tr>
<th>Cloud provider</th>
<th>Link to pricing policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Cloud storage pricing</td>
<td><a href="https://cloud.google.com/storage/pricing">https://cloud.google.com/storage/pricing</a></td>
</tr>
<tr>
<td>IBM SoftLayer pricing</td>
<td><a href="http://www.softlayer.com/info/pricing">http://www.softlayer.com/info/pricing</a></td>
</tr>
</tbody>
</table>

Table 3. Pricing Policy

By limiting the volume of data exiting a cloud provider's data centers, cloud customers are able to control the associated costs. When conducting SAS software cloud deployments, certain design decisions can affect the amount of data transferred, and thus the cost incurred by clients. To illustrate this potential impact, consider a scenario in which SAS software is used to process data sourced from a client's on-premises database (for example, a billing system). Once the data has been processed, the client wants to store the data in a database for analysis by resources who may or may not be SAS software users. There are basically two options available to select from in this scenario, the first being an on-premises database (Figure 2 on the next page).

In Figure 2, the SAS compute server in the cloud pulls data from the on-premises billing system database (labeled as “1”). Other than the small amount of traffic sent from the SAS compute server to the billing system database in order to initiate a connection and send the appropriate code to extract data, data is primarily transferred into the cloud. There is no cost associated with this data transfer. The data is transformed within the SAS compute server and the resulting data is pushed back to the on-premises data warehouse database (labeled as “2”). All data that is pushed to the on-premises data warehouse database accrues a cost for the transfer. Once the SAS compute server has completed the data processing, users within the user zone connect to the data warehouse database using database client software to run extracts and other reports (labeled as “3”). Because this data transfer is wholly within a client’s network, no cloud data transfer costs apply. If users also utilize a SAS software cloud deployment to analyze the data in the data warehouse database, data would be streamed into the cloud for computation and the results would be passed back to the users.
The second option is to have the target database in the same cloud as the SAS software (Figure 3 on the next page). The SAS compute server in the cloud pulls information from the on-premises billing system database (labeled as “1”). Other than the small amount of traffic sent from the SAS compute server to the billing system database in order to initiate a connection and send the appropriate code to extract data, data is primarily transferred into the cloud. There is no cost associated with this data transfer. The data is transformed within the SAS compute server and the resulting data is pushed to the data warehouse database in the cloud (labeled as “2”). Because this data traffic is internal to the cloud, no data transfer costs apply. Once the SAS compute server has completed the data processing, users within the user zone connect to the data warehouse database using database client software to run extracts and other reports (labeled as “3”). Data is returned out of the cloud to the client's network, whether it is generated from SAS clients or from database-specific tools. All of this traffic accrues a cost for the transfer. Data must be in the same availability zone as the SAS software cloud deployment in order to avoid data transfer costs. The determining factors regarding the locality of data are as follows:

- **Regulatory and legal factors** – persisting data in the cloud may not be an option if there are laws or other regulations that prohibit it.
- **Frequency and volume of data updates** – updating changes to customer information is likely to have a low impact, but inserting a large number of transactions or replacing data in a target database can involve a large amount of outbound traffic from the cloud.
- **Types of queries executed against a target database** – selecting counts of customers who have certain attributes is inexpensive from a data transfer cost perspective, but exporting lists of millions of customers and their purchasing histories leads to high costs from a data transfer perspective.
7. General Considerations for SAS®9.4 Software Cloud Deployments

7.1. Selecting a Cloud Provider

Table 4 lists criteria for selecting a cloud provider to host a SAS®9.4 software cloud deployment.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud provider already in place</td>
<td>If cloud provider can meet infrastructure requirements for specific SAS®9.4 software deployment, that provider should be retained; otherwise, cloud provider that can meet requirements at best cost should be found</td>
</tr>
<tr>
<td>Locality of cloud provider</td>
<td>Local regulatory and legal restrictions may affect choice because data is stored or processed in the cloud</td>
</tr>
<tr>
<td>Deploying SAS® Grid Manager</td>
<td>SAS® Grid Manager deployment requires high-performance shared clustered file system, but not all cloud providers offer this capability</td>
</tr>
<tr>
<td>Connectivity options</td>
<td>Cloud customers must understand performance of connections from their premises to the cloud in terms of bandwidth and latency for both data transfers and user experience; security of connection must also be carefully considered</td>
</tr>
<tr>
<td>Costs</td>
<td>Each cloud provider has its own charge model; cloud customers should be diligent in initially assessing anticipated costs and in managing those costs on an on-going basis for their cloud infrastructure</td>
</tr>
</tbody>
</table>

Table 4. Selecting Cloud Provider Criteria
7.2. Connectivity

When considering how a client accesses SAS software cloud deployments, the first item to be considered must be the connection between the client’s network and that of the cloud provider. Security and known performance of the connection is of paramount importance. The recommended best practice is to create a dedicated secure connection that provides sufficient bandwidth with low latency and has service-level agreements in place. Examples of suitable connections are the AWS Direct Connect service and Microsoft Azure’s ExpressRoute. When selecting a cloud provider, clients should ensure that their specific connectivity requirements can be met. When using a secure high-bandwidth connection, the cloud becomes a virtual extension of a client’s network, not unlike having multiple data centers in an existing topology. It is highly recommended for security reasons that SAS software cloud deployments (just as SAS software on-premises deployments) only be accessible from a corporate network, and not from the overall Internet. In general terms, most SAS software cloud deployments can be thought of as consisting of three major zones:

- **User zone:**
  - Segment in which all user personal computers and possibly smart phones and tablets reside
  - Devices are connected to the corporate LAN either via a wired Ethernet connection or a wireless network
  - Typically, wired networks range from 100MbE – 10GbE, with 1GbE currently being the most prevalent
  - Wireless network connectivity, where available, typically uses the 802.11n or 802.11ac standard and provides 450Mbps – 1300Mbps connectivity

- **Corporate server zone:**
  - Segment in which the enterprise-wide infrastructure and services (e.g., Active Directory or other authentication services, operational and data warehouse databases, and big data lakes) reside
  - Network connectivity is typically 10GbE, although some legacy data centers may have 1GbE connectivity
  - Higher-speed InfiniBand networking providing connectivity speeds in excess of 100Gbps may also be used

- **Cloud zone:**
  - Segment in which SAS software and related software are deployed on a cloud infrastructure
  - Network connectivity is 1GbE – 10GbE, but may be limited by the type and size of cloud servers selected.

These three zones and the communication between them are shown in Figure 4 on the next page. The zones are separated from users via firewalls to ensure that access can be controlled. Firewalls are placed at the boundary of each layer (zone). In order for a user’s personal computer to connect to a corporate database, the traffic must be able to go through at least two firewalls, namely one at the exit point of the user zone and another at the entry point of the corporate server zone (labeled as “1”). In order to access a SAS software cloud deployment from a corporate network, users go through the corporate cloud connection to reach the cloud zone (labeled as “2”). Typically, this occurs via a secure dedicated high-bandwidth link between the corporate network and the cloud provider. SAS clients can either be deployed on a bastion host (jump box) running terminal services within the cloud zone, or they can be deployed on user workstations. The SAS infrastructure authenticates users by connecting back to the corporate authentication provider. In order to do this, connectivity must be established from the cloud zone back into the corporate server zone (labeled as “3”).
7.2.1. Site-to-Site Connections

Site-to-site connections (Figure 5 on the next page) create secure shared connections between corporate networks and cloud zones. These connections enable on-premises systems and users to access SAS software in the cloud. Connectivity from the cloud zone back to the corporate server zone may also be enabled, in some cases (for example, to access centralized on-premises authentication providers or data sources and targets). Bi-directional communication between the corporate server zone and the cloud zone is depicted in Figure 4 (labeled as “3”). Site-to-site connections are the only type of connection that allow cloud resources to reach back to corporate resources. However, if a cloud customer does not allow connectivity from the cloud zone back into the corporate server zone, an authentication service could be created in the cloud and data could be pushed from the corporate server zone into the cloud zone. SAS software cloud deployments do not have direct access to the source data on premises. One example of this would be a cloud deployment of Active Directory. Clients can perform a one-way synchronization of Active Directory data into the cloud either for all users and groups or more typically for a subset of users and groups. This approach enables users to log in to SAS software using their Microsoft Windows credentials, without the need for authentication requests to flow back to the corporate server zone.

The data flow for this scenario in which there is no connectivity from the cloud zone into the corporate server zone is shown in Figure 5. In this example, data from the authentication provider within the corporate server zone is pushed into another instance within the cloud (labeled as “1”). Cloud applications are configured to use the cloud-based authentication provider to perform user authentication. Data that must be processed within the SAS software cloud deployment is pushed from on-premises databases in corporate data centers into the cloud (labeled as “2”). This can be accomplished using flat file exports and secure FTP transfers or other database replication mechanisms (assuming that a corresponding replica database was provided in the cloud). SAS application code accesses the data from the cloud. Analysts and end users authenticate themselves via the usual channel (corporate server zone master...
authentication provider) to log into their workstations and access SAS software via the cloud. Log-on requests to the SAS infrastructure would be managed by the cloud authentication provider (labeled as “3”).

Figure 5. Site-to-Site Connection

7.2.2. Point-to-Site Connections

Point-to-site connections (Figure 6 on the next page, with a site-to-site connection labeled as “1”) are created from specific computing devices into the cloud. These connections require the creation of a VPN between Internet-connected computers and the cloud. When users who are not connected to the corporate network require access to the cloud, they must create individual secure connections. This can be achieved by deploying a remote access gateway cloud service and using VPN software on each connecting device (labeled as “2”). Typical use cases for point-to-site connections are as follows:

- Allow resources working from home to access the cloud
- Enable external vendors such as SAS Technical Support to access the cloud
- Enable outsourcing providers to use the SAS infrastructure
- Allow virtual companies without corporate networks to VPN to their cloud resources

While the enablement of point-to-site connections may seem appropriate for these use cases, the same outcome can also be achieved by providing VPN access to the cloud customer’s internal network and leveraging an existing site-to-site connection for reaching into the cloud. Point-to-site connections are a great choice when a cloud customer does not have a site-to-site connection in place and wishes to conduct a pilot cloud deployment without committing the time and financial investment required to implement a site-to-site connection.
### 7.2.3. Static versus Dynamic IP Addresses and Host Names

The default behavior when starting a cloud server instance is that the instance is allocated a dynamic private IP address. Some cloud providers such as AWS create host names from this IP address (for example, an instance that has a private IP address of 172.31.48.213 and a private domain name server [DNS] name of ip-31-48-213.ec2.internal). Upon a server re-start, the IP address can change, which in turn generates a new host name. This creates a challenge in SAS software deployments because it is configured for specific host names (DNS aliases) during installation. It is highly likely that a change in a server’s host name would cause issues with the operation of SAS software. To avoid this situation, we recommend that clients allocate static IP addresses to their SAS server instances and configure the instances to use specific fully qualified DNS names. This can be achieved as follows:

1. Create a virtual network interface (elastic network interface [ENI] in AWS)
2. Allocate a fixed IP address from the range of IP addresses for the subnet where the instance is created to the ENI.
3. Define a new instance by selecting the ENI that was created rather than by allowing the instance to create a default ENI.
4. Configure the instance to use a fixed host name.
5. Add the DNS names and IP address to the corporate DNS.
6. Configure the instances to interface with the corporate DNS.

The end result of this configuration is that instances have static IP addresses and fixed DNS names, which ensures that SAS software functions subsequent to a re-start of an instance. It is critical that this be in place prior to the SAS software deployment so that the static DNS alias of the host name is embedded in the SAS software configuration.
7.2.4. Outbound Networking Access

SAS Visual Analytics requires access to Open Street Map servers that we host in order to access mapping tiles. SAS Visual Analytics cloud deployments must be able to access these servers, so cloud servers require Internet access to them. In order to provide Internet access from the private subnets where SAS Visual Analytics is deployed, it is necessary to do the following:

1. Create a public subnet in the VPC.
2. Add either a network address translation (NAT) gateway or a NAT instance in the public subnet.
3. Allocate an elastic IP address to the NAT gateway (a requirement of AWS).
4. Add a rule in the private subnet's routing table to direct Internet traffic to the NAT gateway.
5. Add a rule in the public subnet's routing table to direct Internet traffic to the Internet gateway.

A NAT gateway ensures that SAS Visual Analytics servers in the private subnet can access the Internet while simultaneously blocking Internet-sourced connections from creating connections with SAS Visual Analytics servers. The private subnet's routing table should be updated as shown in Table 5, assuming that the cloud subnet is defined as 10.0.0.0/16 in classless inter-domain routing (CIDR) notation.

<table>
<thead>
<tr>
<th>Destination (CIDR reference)</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/16</td>
<td>Local</td>
</tr>
<tr>
<td>149.173.160.5/32</td>
<td>NAT gateway ID</td>
</tr>
</tbody>
</table>

*Table 5. Private Subnet Settings*

The public subnet's routing table should be updated as shown in Table 6, assuming that the cloud subnet is defined as 10.0.0.0/16 in CIDR notation.

<table>
<thead>
<tr>
<th>Destination (CIDR reference)</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0.0.0/16</td>
<td>Local</td>
</tr>
<tr>
<td>149.173.160.5/32</td>
<td>Internet gateway ID</td>
</tr>
</tbody>
</table>

*Table 6. Public Subnet Settings*

In this example, IP address 149.173.160.5 is the virtual IP address of Open Street Map servers *opnsta.sas.com*, *opmstb.sas.com*, and *opnstc.sas.com*. The NAT gateway ID should be replaced with the ID of the NAT gateway instance, and the Internet gateway ID should be replaced with the ID of the Internet gateway instance. Other SAS software can be used to access and consume Web and other services over the public Internet. If a client makes use of these capabilities, additional corresponding entries would be required in the routing tables.

7.2.5. IP Multicast Support

IP multicasting is no longer used on the SAS®9.4 platform by default. However, if a client has created custom applications that have been or that will be migrated to the SAS®9.4 platform in earlier releases of SAS software, and if those custom applications make use of the SAS remote services application, it is likely that these applications would rely on multicasting. AWS, Google Cloud, IBM SoftLayer and Microsoft Azure all state that IPv4 multicasting is not supported in their cloud infrastructures. If a client has built custom SAS software that uses multicasting, such deployments would not work in the cloud and must be re-built if a SAS software cloud deployment is desired.
7.3. Impact of SAS Client Locations

Communication between SAS thick clients and SAS servers generates varying amounts of data transfers. Due care should be taken to determine the most appropriate location for installing SAS clients during the design stage. Table 7 provides available options with their pros and cons.

<table>
<thead>
<tr>
<th>Location of SAS clients</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jump box in the cloud</td>
<td>Access is via terminal services (remote desktop connections)</td>
<td>Requires licensing Microsoft® Terminal Services or similar services</td>
</tr>
<tr>
<td></td>
<td>Only data that describes screens is transferred</td>
<td>Can be confusing for users who are not used to working on remote desktop connections</td>
</tr>
<tr>
<td></td>
<td>Upgrade and patching of client software can be accomplished centrally, and there is no requirement to create and manage deployment packages</td>
<td>Drive letters and network resources that users are used to will more than likely either be different or be totally unavailable</td>
</tr>
<tr>
<td>On-premises; installed directly on user workstations</td>
<td>Simpler for users Users work on their own personal computers, so the cons of jump boxes in the cloud are avoided</td>
<td>More data is transferred from the cloud to user workstations, especially when viewing data tables or exporting to locations within corporate networks For SAS®9.4 software that interacts heavily with SAS® Metadata Server, speed of connections as well as latency can materially affect how responsive SAS®9.4 software appears to end users Clients typically do not allow users to have ability to install and upgrade software, so SAS clients are commonly packaged and deployed by client IT teams, which makes it more difficult to keep them up to date</td>
</tr>
</tbody>
</table>

Table 7. SAS Client Locations

Hybrid solutions could also be implemented such that there would be some applications installed on premises and others that would be in the cloud. The decision comes down to looking at the pros and cons and weighing the costs and impacts of each option.

8. Scalability in the Cloud

One of the benefits cloud computing provides to clients is the ability to scale in an agile manner. By continually sizing cloud computing to meet current demand, clients can manage their cloud infrastructure costs. As workload increases, additional computing power can be provided. There are several methods by which a cloud can scale, and these methods are discussed in this section. Once the various methods of scalability have been defined, the applicability of each method is discussed in relation to the various SAS server types.

8.1. Scale Up

Scaling up (Figure 7 on the next page) is the process of making an existing server larger by adding more CPU and/or RAM and potentially also by increasing the I/O capabilities and network connectivity speed. As usage on applications
increases, cloud servers can be re-sized to meet demand such that a server could be changed from a 4-vCPU server to an 8-vCPU server with additional RAM to meet the higher workload. Increasing the size of a server or servers enables greater processing capacity, but also incurs a higher cost from the cloud provider. Additionally, scaling up a server or servers requires that those servers be re-started. During the time that the servers are re-starting, the services provided by those servers are unavailable.

**Figure 7. Scaling Up**

### 8.2. Scale Down

Scaling down (Figure 8) is the process of making an existing server smaller by reducing CPU and/or RAM and potentially also by decreasing the I/O capabilities and network connectivity speed. During periods of lower usage of applications, cloud servers can be re-sized to lower specifications such that a server could be changed from an 8-vCPU server to a 4-vCPU server whenever usage is low. Decreasing the size of a server or servers reduces the processing capacity, but also incurs a lower cost from the cloud provider. Additionally, scaling down a server or servers requires that those servers be re-started. During the time that the servers are re-starting, the services provided by those servers are unavailable.

**Figure 8. Scaling Down**

### 8.3. Scale Out

Scaling out (Figure 9 on the following page) is the process of adding extra servers to share processing load in order to provide higher processing capacity. In this example, an additional SAS middle-tier server and SAS compute server were added to the SAS® 9.4 software deployment in order to handle additional load.
8.4. Scale In

Scaling in (Figure 10) is the process of removing one or more servers in order to provide lower processing capacity. In this example, one of the SAS middle-tier servers was no longer needed as load decreased and was switched off. Later, the load on the SAS compute servers was also reduced to a point where only one SAS compute server was required. Thus, one of the SAS compute servers was also switched off.

Figure 10. Scaling In
8.5. Elasticity

An elastic application or infrastructure is one that can automatically scale out and in dependent upon the load that the application is under. Having elasticity has the benefits of maximizing performance for the user community and batch processing, as well as the ability to manage costs by reducing the size of the SAS infrastructure when demand is lower.

8.6. Implications of Scaling SAS®9.4 Software

The potential implications of SAS software availability and licensing must be considered prior to scaling a SAS software cloud deployment. A number of icons are used are used in the following subsections. The meanings of those icons are shown in Table 8.

<table>
<thead>
<tr>
<th>Icon</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>No re-configuration is necessary and no outage is required; however, any connections to service and associated user sessions are dropped, in which case users must re-connect and log in again in order to continue working</td>
</tr>
<tr>
<td>!</td>
<td>Re-configuration <em>may</em> be required</td>
</tr>
<tr>
<td>!</td>
<td>Re-start of service is required, which causes outage of SAS infrastructure</td>
</tr>
<tr>
<td>✗</td>
<td>Invalid action or option</td>
</tr>
</tbody>
</table>

Table 8. Icon Meanings

8.6.1. SAS® Metadata Server

SAS Metadata Server can be deployed on either a single host or in a cluster of three or more hosts. The various options for scaling SAS Metadata Server are shown in Table 9. For SAS software deployments that do not include SAS Grid Manager, there is no restriction on the size or number of SAS Metadata Servers as long as the servers running SAS Metadata Server processes are *not* also being used for SAS compute processing. The licensing requirements for a SAS® Grid Manager cloud deployment are exactly the same as for a SAS software on-premises deployment.

<table>
<thead>
<tr>
<th>Scaling approach</th>
<th>Single SAS® Metadata Server</th>
<th>Clustered SAS® Metadata Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale up</td>
<td>! Requires re-start of entire SAS infrastructure</td>
<td>✓ If servers are taken off-line one at a time, additional capacity is allocated and re-started and there are no service outages</td>
</tr>
<tr>
<td>Scale down</td>
<td>! Requires re-start of entire SAS infrastructure after SAS® Metadata Server is re-started</td>
<td>✓ If servers are taken off-line one at a time, capacity is reduced and re-started and there are no service outages</td>
</tr>
</tbody>
</table>
Scaling approach | Single SAS® Metadata Server | Clustered SAS® Metadata Server
--- | --- | ---
Scale out | ![⚠️] If extra servers are added for metadata, they must form cluster and metadata cluster quorum rules would apply; this would require SAS infrastructure outage and changes to configuration in order to support clustering | ![✅] Additional servers can be added to metadata cluster subject to quorum rules; this would require re-configuration in order to include new hosts
Scale in | ![❌] Not applicable; SAS®9.4 software deployments require running SAS® Metadata Server; if single SAS® Metadata Server is stopped, entire SAS infrastructure would not work | ![✅] No service outages are required as long as there are two or more SAS® Metadata Servers in cluster; if exactly two SAS® Metadata Servers are running, failure of one host would cause SAS infrastructure outage

Table 9. SAS® Metadata Server Deployment

8.6.2. SAS Web Servers

SAS Web servers can be deployed on either a single host or in a cluster of two or more hosts. The various options for scaling SAS Web servers are shown in Table 10. For SAS software deployments that do not include SAS Grid Manager, there is no restriction on the size or number of SAS Web servers as long as the servers running SAS Web server processes are not also being used for SAS compute processing. The licensing requirements for a SAS® Grid Manager cloud deployment are exactly the same as for a SAS software on-premises deployment.

<table>
<thead>
<tr>
<th>Scaling approach</th>
<th>Single SAS Web server</th>
<th>Clustered SAS Web server</th>
</tr>
</thead>
</table>
| Scale up | ![⚠️] Requires re-start of SAS Web server; Web applications are not accessible during re-starts | ![✅] If servers are taken off-line one at a time, additional capacity is allocated and re-started and there are no service outages

Scale down | ![⚠️] Requires re-start of SAS Web server; Web applications are not accessible during re-starts | ![✅] If servers are taken off-line one at a time, capacity is reduced and re-started and there are no service outages

Scale out | ![✅] If extra servers are added, external load balancer is required to direct traffic between SAS Web server hosts | ![⚠️] External load balancer must be re-configured to include new hosts

Scale in | ![❌] Not applicable; SAS®9.4 software deployments require running SAS Web server; if single SAS Web server is stopped, SAS Web applications would not be accessible | ![✅] No service outages are required as long as at least one SAS Web server remains running

Table 10. SAS Web Server Deployment

8.6.3. SAS Web Application Servers

SAS Web application servers can be deployed on either a single host or in a cluster of two or more hosts. The various options for scaling SAS Web application servers are shown in Table 11. For SAS software deployments that
do not include SAS Grid Manager, there is no restriction on the size or number of SAS Web application servers as long as the servers running SAS Web application server processes are not also being used for SAS compute processing. The licensing requirements for a SAS® Grid Manager cloud deployment are exactly the same as for a SAS software on-premises deployment.

<table>
<thead>
<tr>
<th>Scaling approach</th>
<th>Single SAS Web application server</th>
<th>Clustered SAS Web application server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale up</td>
<td>⚠️ Requires re-start of SAS Web application server; Web applications are not accessible during re-starts</td>
<td>✓ If servers are taken off-line one at a time, additional capacity is allocated and re-started and there are no service outages</td>
</tr>
<tr>
<td>Scale down</td>
<td>⚠️ Requires re-start of SAS Web application server; Web applications are not accessible during re-starts</td>
<td>✓ If servers are taken off-line one at a time, capacity is reduced and re-started and there are no service outages</td>
</tr>
<tr>
<td>Scale out</td>
<td>✓ If extra servers are added, SAS Web server must be re-configured to perform load balancing of SAS Web applications</td>
<td>✓ If extra servers are added, SAS Web server must be re-configured to perform load balancing of SAS Web applications</td>
</tr>
<tr>
<td>Scale in</td>
<td>✗ Not applicable; SAS®9.4 software deployments require running SAS Web application server; if single SAS Web application server is stopped, SAS Web applications would not be accessible</td>
<td>✓ No service outages are required as long as at least one SAS Web application server remains running</td>
</tr>
</tbody>
</table>

Table 11. SAS Web Application Server Deployment

8.6.4. SAS Compute Server

The SAS compute server tier (Table 12) is the tier on which SAS programs run in order to process data. For the majority of SAS software, the SAS compute server is restricted in the number of CPUs that can be allocated to it by the license agreement.

<table>
<thead>
<tr>
<th>Scaling approach</th>
<th>SAS compute server without SAS® Grid Manager</th>
<th>SAS compute server with SAS® Grid Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale up</td>
<td>⚠️ Requires re-start of SAS compute server; data processing is not possible during re-starts</td>
<td>✓ If servers are taken off-line one at a time, additional capacity is allocated and re-started and there are no service outages, but processing capacity is reduced until all servers are back on-line</td>
</tr>
<tr>
<td>Scale down</td>
<td>⚠️ Requires re-start of SAS compute server; data processing is not possible during re-starts</td>
<td>✓ If servers are taken off-line one at a time, additional capacity is allocated and re-started and there are no service outages, but processing capacity is reduced until all servers are back on-line</td>
</tr>
</tbody>
</table>
Table 12: SAS Compute Server Deployment

Scaling out the SAS compute server tier requires that there is a high-performance shared file system available for all SAS grid compute servers. This capability is currently only available from some cloud providers, but not all of them.

8.6.5. SAS® Visual Analytics LASR™ Nodes

LASR™, the in-memory engine behind SAS Visual Analytics, can be deployed in one of two manners, namely on a single machine (non-distributed) or across multiple machines (distributed). LASR nodes are licensed according to the number of CPU cores across which it is being deployed. Therefore, if a client must scale its SAS Visual Analytics deployment up or out, it must ensure that it remains in line with the terms of the license. This is shown in Table 13.

Table 13: SAS® Visual Analytics LASR™ Nodes Deployment
For distributed SAS Visual Analytics, cluster performance is limited by the node with the least capacity and performance assuming that data is uniformly distributed across all nodes. Thus, it is important that all nodes have the same allocation of CPU, memory, storage, and network resources when scaling a cluster up or out. Furthermore, if servers are added or removed, it may be necessary to re-stage the data in HDFS (Hadoop Distributed File System) in order to ensure that all servers have an equal distribution of data and that all slices of data have the intended number of replicas. Refer to the SAS Visual Analytics administration guide for more information regarding data distribution.

9. Architectural Best Practices

9.1. Authentication

Authentication is the process in which a user requesting access to an application has their credentials verified.

9.1.1. Administrative Access to Cloud Servers

Administrators often require access to a server that is part of the SAS®9.4 software deployment. For Microsoft Windows servers, such access can be provided using remote desktop connections and is controlled in exactly the same way as for a SAS software on-premises deployment. When the SAS servers are on the Linux operating system, however, the default access method requires the use of a public or private key exchange in order to enable log-in. This is different than it is for SAS software on-premises deployments, where it is more common for user IDs and passwords to be used for log-ins. If a client requires user IDs and passwords rather than a public or private key, additional configuration of the Linux servers in the cloud would be required.

9.1.2. SAS Application Authentication

Except in the special case of internal metadata accounts for SAS software-related services, SAS software relies on external authentication of users. This is done either at the operating system on which the SAS software is running or via an external identity management service using the LDAP protocol. In the event that direct access to an LDAP-compliant service is used, this can also be performed securely using the LDAPS protocol. For SAS software deployments on the Microsoft Windows operating system, the operating system is typically integrated with a centralized authentication service such as Active Directory. For SAS software deployments on the UNIX and Linux operating systems, local host accounts and groups were once the norm, but it is becoming increasingly common for clients to use a PAM configuration that integrates UNIX and Linux servers with Active Directory and other centralized authentication services. The benefits of binding the operating system to a centralized authentication provider are as follows:

• IT benefits:
  – There is a single point of control for the management of users and groups.
  – Support effort is reduced because there is no requirement to map different credentials across operating systems.
  – SAS software installations are simplified because, as far as SAS software is concerned, authentication is being handled by the underlying host operating system.

• User benefits:
  – Users can access SAS applications using a single set of credentials (e.g., Microsoft Windows credentials) regardless of the platform that SAS software is running on.
  – Users do not have to remember user IDs and passwords for multiple systems.
There are two options available to clients wishing to leverage centralized authentication for SAS software cloud deployments:

• Configure connectivity between the cloud servers to the on-premises authentication service:
  – This requires changes in firewall rules in the client’s corporate server zone.
  – The client must ensure that inbound access to services is accomplished over a secure connection and that access is restricted to known IP address ranges of the cloud servers.

• Deploy another authentication service within the cloud and synchronize appropriate content between the on-premises authentication service and the corresponding cloud-hosted service:
  – The cloud servers on which SAS software is installed would be configured to use the cloud authentication service, and no direct access would be required from the cloud servers back into the on-premises authentication service.

9.2. Create VPC

Cloud providers allow cloud customers to deploy virtual servers and other services on a shared infrastructure. Because cloud servers are running “next to” those of other cloud customers, the ability to ring-fence servers from all others is vital. This can be achieved by creating a private network (for example, AWS VPCs and Microsoft Azure virtual networks) in the cloud. VPCs support connecting existing data centers to cloud resources using industry-standard IPsec VPN connections. One or more subnets can be created within a VPC and allow separation of the various layers of cloud services at the networking layer. At an absolute minimum, the following subnets should be created:

• SAS client subnet – contains servers that have SAS clients installed on them
• SAS server subnet – contains servers running SAS servers
• Demilitarized zone (DMZ) subnet – public-facing servers.

In order to align with common practices for client IT standards, a more granular approach to the creation of subnets is recommended:

• SAS client subnet – contains servers that have SAS clients installed on them.
• SAS Metadata Server subnet – contains SAS Metadata Servers
• SAS middle-tier subnet – contains SAS Web server and SAS Web application servers; designed to be internally facing from the client's perspective
• SAS compute tier subnet – contains SAS compute servers and may have access to databases in the client’s internal network
• SAS database subnet – contains the databases accessed by the SAS compute tier
• DMZ subnet – public-facing servers.

In order to enable connectivity between the different subnets, cloud customers must define firewall rules just as they would for their on-premises data centers. Some cloud providers have different names for firewall rules (for example, they are known as security groups in AWS).

9.3. Server Locations

Cloud providers have multiple data centers (for example, AWS availability zones and Microsoft Azure availability sets) in which cloud customers can deploy cloud services. Some geographic locations can have multiple data centers. In order to avoid data transfer costs, clients must ensure that all SAS servers, storage, and associated cloud infrastructure are stood up in the same region. SAS software can be deployed as shown in Table 14.
Option | Comments
--- | ---
Wholly within single availability zone or availability set | If availability zone or availability set becomes unavailable, there is outage of SAS infrastructure
Deployed in active–passive manner across availability zones or availability sets | SAS®9.4 software access and processing can be recovered by starting SAS®9.4 software in second availability zone or availability set; this is typical D/R approach in production

Table 14. SAS®9.4 Software Deployment Server Locations across Availability Zones/Availability Sets

9.4. HA

In a SAS software on-premises deployment, HA is provided through the following:

- Deploying SAS services on multiple servers in an active–passive manner
- Clustering SAS services across two or more servers
- Combination of these.

For SAS software cloud deployments, the approach varies slightly because clients do not have control over which physical hosts run their cloud servers within an availability zone or availability set. By using different availability zones or availability sets within the same cloud region, clients can ensure that services are actually running on separate physical hosts. Care must be taken when deploying SAS software in multiple availability zones or availability sets, however, because network bandwidth between availability zones and availability sets is lower than network bandwidth within a single availability zone or availability set. This can have performance implications for SAS software, especially on a SAS grid infrastructure where SAS compute nodes require high network bandwidth to the shared clustered file system (CFS).

9.5. D/R

In SAS software on-premises deployments, a separate D/R environment is typically deployed into a separate data center than that in which the PROD environment is deployed. By using different data centers, clients are able to recover their SAS processing capabilities in the event of a failure of the primary site. For SAS software cloud deployments, it is recommended that separate cloud regions be used for PROD and D/R environments. While it is extremely unlikely that a cloud provider's entire region would go off-line, it is possible due to natural disasters or failure of a main internet connection. Clients must be aware that in order for the production data and configuration to be synchronized to the D/R zone, the transfer of data between zones incurs data transfer fees. The geographic location of zones must also be taken into account, and it is recommended that the zones used for both the PROD and D/R environments be within the same country in order to ensure compliance with legislation pertaining to data storage.

9.6. Maximizing Network Throughput within the Cloud with AWS Placement Groups

While extremely low latency and maximum throughput at the network layer are required, SAS recommends that clients consider using an AWS placement group and “place” all of a cloud infrastructure for applications that require
such connectivity within the same group. Two caveats that may make using groups inappropriate for a client are as follows:

- Cloud providers recommend that all instances within a group should be of the same instance type and support 10GbE networking.
- A request to re-start a single on-demand instance that is in a group cannot be satisfied; all instances within a group must be re-started, which causes an outage of service for the SAS infrastructure.

### 9.7. Storage Selection

During the process of creating a server in the cloud, there are a number of options regarding the provisioning of storage. While it is possible to define a server with a number of disks for storing application software, configuration details, and even data, this storage is tied to a server and the server therefore becomes stateful (that is, the data and server are inherently tied together). A recommended design principle is to make servers stateless, where a server and its local storage are managed separately from the storage where application software, configuration details, and data are stored. A stateless server has the distinct advantage that it can be more easily replaced with another server, and once created storage can be mounted on to it. SAS software deployments require three or four types of storage depending upon the SAS software being deployed:

- **Server storage** – can be provisioned using traditional mechanical hard drives, solid state drives (SSDs), and high-performance SSDs with higher IOPS ratings; can only be mounted or attached to one server at a time; known as elastic block storage (EBS) in AWS
- **Shared storage** – Intel Lustre provides a POSIX-compliant file system that can be mounted to multiple servers concurrently; IBM offers their elastic storage (GPFS) in its SoftLayer cloud
- **Ephemeral storage** – similar to single-server storage, but with the major exception that the file system is not persisted between re-starts; thus, it is not suited to storing application software, configuration details, or permanent data
- **Object storage** – provides highly redundant storage at a competitive price point; known as S3 storage in AWS.

We recommend the storage types shown in Table 15.

<table>
<thead>
<tr>
<th>Storage type</th>
<th>Use in SAS®9.4 non-grid deployments</th>
<th>Use in SAS®9.4 grid deployments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server storage</td>
<td>SAS®9.4 software; SAS®9.4 software configurations; permanent data</td>
<td>Not used</td>
</tr>
<tr>
<td>Shared storage</td>
<td>Not used</td>
<td>SAS®9.4 software; SAS®9.4 software configurations; permanent data</td>
</tr>
<tr>
<td>Ephemeral storage</td>
<td>SASWORK and UTILLOC</td>
<td>SASWORK and UTILLOC</td>
</tr>
<tr>
<td>Object storage</td>
<td>Snapshots (back-ups) of server storage</td>
<td>Snapshots (back-ups) of server storage</td>
</tr>
</tbody>
</table>

*Table 15. Storage Types*

### 9.8. Grid-Specific Information

If the SAS software includes SAS Grid Manager, there are some extra considerations that must be accounted for. The first is to ensure that the SAS grid infrastructure has access to a high-bandwidth shared CFS. We have performed testing on the Intel Lustre file system in an AWS cloud deployment and have found it to be suitable as a cloud-based shared CFS. Further information regarding our testing can be found on the SAS Support Web site located at [https://support.sas.com/rnd/scalability/grid/SGMonAWS.pdf](https://support.sas.com/rnd/scalability/grid/SGMonAWS.pdf).
If an IBM SoftLayer cloud is being used, IBM elastic storage (previously known as GPFS) is available over 10GbE-network connections. In a PROD environment, the SAS compute nodes within the SAS grid infrastructure should be built with support for 10GbE networking, and it is recommended that jumbo frames be enabled on these hosts in order to maximize data transfers. Ideally, SAS compute nodes should have dual 10GbE-network interfaces, namely one for general purposes and the other dedicated for accessing the shared CFS. Not all instance types offered by cloud providers have the ability to support 10GbE networking and jumbo frames.

If a client chooses to build a SAS grid infrastructure from more, smaller instances that only support 1GbE networking, it is recommended that the performance of the selected instance types within the cloud be benchmarked and that there be an awareness of the fact that overall performance of the solution may be limited by network I/O throughput. For SAS metadata, SAS Web servers, and SAS Web application servers, as well as for non-PROD environments in which performance is of less importance, the use of 1GbE-network interfaces is sufficient.

### 9.9. Monitoring and Auditing Cloud Deployments

In general, monitoring and auditing principles for a SAS software cloud deployment are the same as for a SAS software on-premises deployment. Details can be found at [Logging and Monitoring for SAS Servers](#). In addition, monitoring the instances themselves should be considered, and any VPN connections should also be monitored.

#### 9.9.1. Monitoring Instances

Monitoring instances to ensure that they are up and running can be accomplished via any third-party tool already being used in on-premises circumstances. However, the command and control server of the monitoring tool should not be placed within monitored instances. Therefore, it is not recommended that SAS® Environment Manager, which provides basic machine monitoring capabilities, be the sole monitoring solution for cloud deployments. Currently, SAS Environment Manager is not able to operate in an HA mode, and it would be a single point of failure. Because SAS Environment Manager server components are a component of the SAS middle tier, the ability of SAS Environment Manager to provide insight is dependent on the availability of the SAS middle tier. Monitoring facilities provided by the cloud provider of choice (for example, Amazon CloudWatch and Microsoft Azure Portal Monitor) can also be used for these monitoring purposes.

#### 9.9.2. Monitoring VPN Connections

If VPN connections such as site-to-site connections or point-to-site connections are being used, they must also be monitored. If a VPN connection fails, SAS software cloud deployments appear from the client perspective as being down, although in reality they may only be inaccessible due to a failed VPN connection.

### 10. Backing Up and Restoring SAS®9.4 Software in Cloud Infrastructures

Whether SAS®9.4 software is running on premises or in the cloud hosted via an IaaS provider, the assets that form part of the SAS infrastructure must fall under a comprehensive back-up strategy in order to facilitate business objectives in the areas of continuity and compliance. In addition, clients must also ensure that they have a tried-and-tested process to recover their SAS infrastructure and underlying cloud infrastructure. The only way a client can prepare documentation for the recovery process tailored to their SAS infrastructure is to execute a restore. The best time to do this is prior to the go-live date for a PROD environment. Having a clone of the PROD environment in which such testing can occur also helps in this regard. The same applies to SAS software on-premises deployments.
10.1. What to Back Up

10.1.1. Servers

VM instances that comprise the server tier of the SAS infrastructure should ideally be backed up after the original installation and configuration in order to provide a baseline starting point. Images can be created from configured servers to create the baseline from which new instances can be provisioned, either serving as back-up replacements or being ready to stand up new instances in different environments. The process of working with baseline images shortens the time to recovery for failures or provision new environments considerably. When coupled with automation techniques, this process can be at least partially (if not fully) automated to create systems that can self-heal in order to achieve HA or push-button deployment of new environments, thus shortening the recovery time objective in the D/R strategy.

10.1.2. Content and Data

Data and content created by and consumed by the SAS infrastructure must be backed up in conjunction with the SAS metadata, SAS® Content Server, SAS® Web Infrastructure Platform Database, and other platform database repositories because they are dependent on one another and must be kept in sync. The physical content that must be part of the back-up strategy depends on the SAS software being deployed. Some examples of content that should be backed up include the following:

- Source code (e.g., SAS® Stored Processes, SAS analytical models, and SAS® Data Integration Studio jobs)
- SAS data files
- WebDAV content (e.g., reports and documents)
- Relational databases.

10.2. Backing Up

All data that must be archived for back-up and restoration purposes should be stored on suitable disk volumes. A strategy for creating regular snapshots from these volumes and required retention periods should be defined. Snapshots are incremental point-in-time back-ups of volumes stored in lower-cost object storage, where they are redundantly replicated across multiple availability zones or availability sets. Because snapshots are block-based, space is consumed (and therefore costs are incurred) only for changed data after the initial snapshot has been created.

10.2.1. Cross-Region Back-Ups

Snapshots can be copied across cloud provider regions to provide an even higher level of redundancy and guard against region failure.

10.2.2. Consistent Back-Ups

Ideally the SAS infrastructure should not be performing any I/O when back-ups occur. If possible, the file system should be quiesced prior to attempting snapshot creation. If the file system or volume manager in use does not support the ability to freeze I/O, one option is to unmount the file system, perform the snapshot, and then re-mount the file system. Because the snapshot process is quick to execute and captures a point in time, volumes only have to be unmounted for a matter of seconds. Thus, the back-up window is small and can be scheduled effectively.
10.2.3. Multi-Volume Back-Ups

If striping is used across multiple EBS volumes to increase performance using a logical volume manager, back-ups should be performed from the volume manager and not directly from the individual storage volumes. Traditional agent-based enterprise back-up solutions can be utilized to accomplish consistent back-ups of striped volumes. Alternatively, the target location for back-up storage from these solutions could be a large disk volume that can be subsequently backed up via snapshots. This depends on the amount of data to be backed up and whether or not a large enough volume to hold the back-up data can perform well.

10.3. Restoring

In order to restore data from a snapshot, a new volume is created from it. The existing volume is unmounted and detached from the instance and the new volume is attached and mounted. Alternatively, a new instance could be launched from the snapshot and the existing instance could be entirely replaced. It is also possible to perform a partial restore by attaching the new volume created from the snapshot to the instance, mounting it at a different mount point, and copying the data from the new volume to the existing volume utilizing operating system commands. The new volume would then be unmounted, detached, and deleted.

10.4. Automation

Cloud providers provide several APIs and tools that can assist with automating back-up and restore processes.

10.4.1. On-Instance Scripts

Shell scripts such as PowerShell for Microsoft Windows that automate back-up, restore, attaching and detaching, and mounting and unmounting of volumes can be deployed on instances and can be run via operating system schedulers (for example, cron and Microsoft Windows Task Scheduler).

10.4.2. AWS Lambda

AWS Lambda can be utilized to perform scheduled automated back-ups of the entire set of instances and manage the retention of snapshots. Logic can be programmed to take advantage of volume and instance tagging in order to determine what to back up and retention policies to be applied, along with possible cross-region copying.

10.4.3. Third-Party Solutions

Cloud providers offer add-on solutions for automation. The AWS Marketplace and the Microsoft Azure Marketplace provide access to many vendor solutions in the area of storage and back-up management.

11. Resources

The following contain materials that provide further information regarding cloud computing in relation to SAS®9.4 software:

- AWS Placement Groups
- CIDR (Wikipedia)
• Designating Ports and Multicast Addresses
• Network Maximum Transmission Unit (MTU) for Your EC2 Instance (especially Jumbo Frames (9001 MTU))
• Performance and Tuning Considerations for SAS® Grid® Manager 9.4 on Amazon (AWS) Cloud using Intel Cloud Edition for Lustre File System
• The NIST Definition of Cloud Computing.
To contact your local SAS office, please visit: sas.com/offices