Best Practices for Deploying SAS Server on AWS

February 2020

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Contents

About this Guide

Many SAS customers are moving their SAS applications from on-premises data centers to the AWS Cloud. In order to migrate, customers must be aware of all the layers of their SAS infrastructure. Customers should understand how their SAS applications run, and how to optimize their Amazon Web Services (AWS) architecture.

This whitepaper addresses performance considerations and best practices for SAS®9 (SAS® Foundation and SAS Grid Manager) and SAS® Viya® when hosted on AWS. The content is written for IT professionals familiar with SAS and AWS.

Introduction

SAS is an analytics software that provides organizations a suite of capabilities that enable users to draw insights from data and make intelligent decisions. The SAS platform includes software platforms that underpin SAS product offerings in analytics, data management, and visualization. SAS 9.4 provides simplified architecture and deployment options for running SAS on a cloud infrastructure. SAS Viya is a cloud-enabled, in-memory analytics engine that provides quick, accurate and reliable analytical insights.

SAS 9.4 provides the following features:

1. Data Management
2. Visual Analytics
3. Governance and Security
4. Forecasting and Text Mining
5. Statistical Analysis
6. Environment Management

SAS is also a 4GL programming language used by data scientists for more than 80,000 customers globally.

SAS Architecture

SAS 9.4 Intelligence Platform

This platform is designed to efficiently access large amounts of data, while simultaneously providing timely intelligence to a large number of users. The platform uses an n-tier architecture that allows you to distribute functionality across compute resources, so that each type of work is performed by the resources most suitable for the job.

This architecture consists of the following four tiers:

- Data sources that store customers enterprise data
- SAS servers that perform SAS processing on enterprise data

- Middle tier that enables users to access intelligence data and functionality through a web browser, and provides shared services used by the platform's applications
- Client tier that provides desktop access to intelligence data and functionality through easy-to-use interfaces

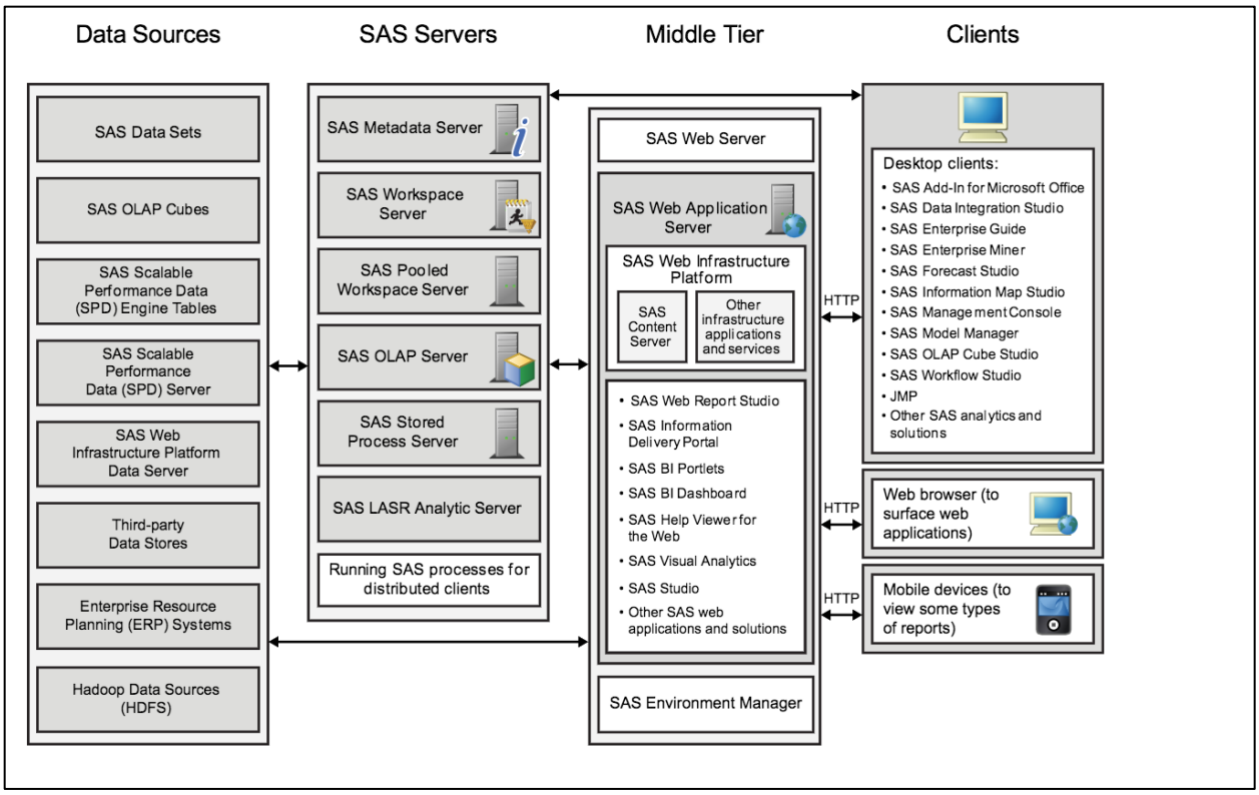

Figure 1: SAS 9.4 Intelligence Platform Architecture

SAS 9.4 Grid

The SAS grid computing environment uses SAS Grid Manager to distribute SAS computing tasks among multiple computers on a network. Workload distribution enables the following functionality:

- Workload balancing allows multiple users in a SAS environment to distribute workloads to a shared pool of resources.
- Accelerated processing allows users to distribute subtasks of individual SAS jobs to a shared pool of resources.
- Scheduling jobs allows users to schedule automatically routed tasks to the shared resource pool.

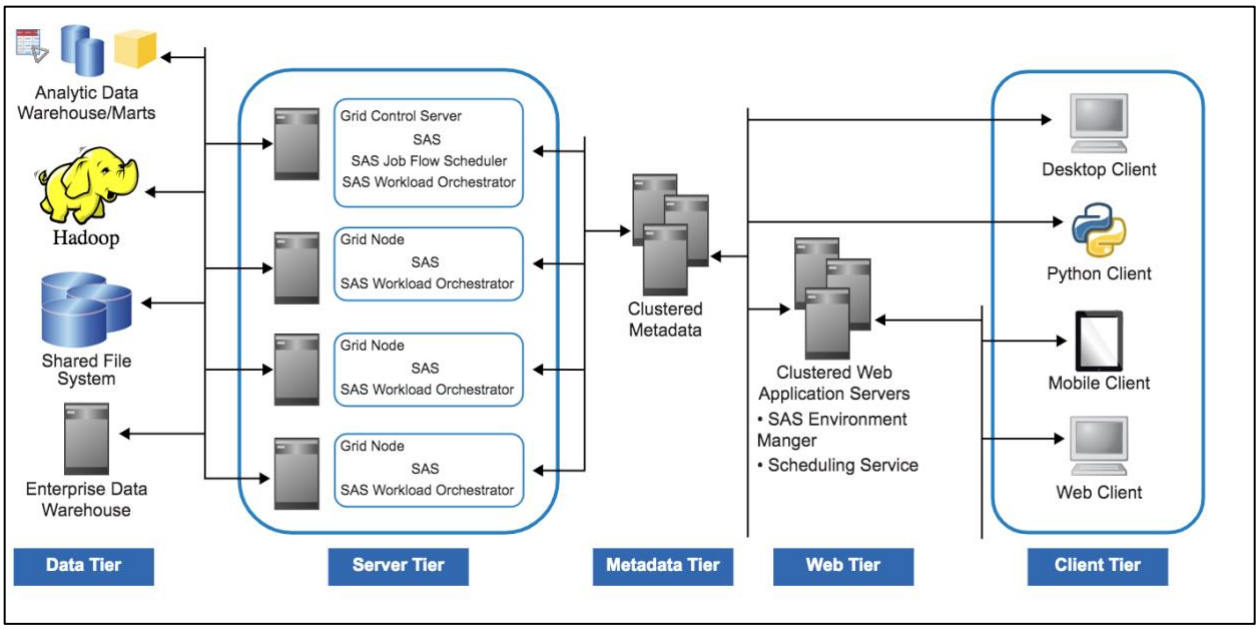

Figure 2: SAS Grid Architecture

SAS Viya

SAS Viya is a cloud-enabled, in-memory analytics engine that provides quick, accurate and reliable analytical insights. Elastic, scalable, and fault-tolerant processing addresses complex analytical challenges while effortlessly scaling for future use cases. SAS Viya has the following benefits:

- SAS Viya provides distributed analytical in-memory calculations that are optimized for unconstrained environments and automatically adjust in constrained environments.
- SAS Viya supports a standardized code base that enables programming in SAS and other languages like Python, R, Java, and Lua.
- SAS Viya is highly available with distributed processing crafted to handle multiple users distributing operations across the cores of a single server, or nodes of massive compute clusters.

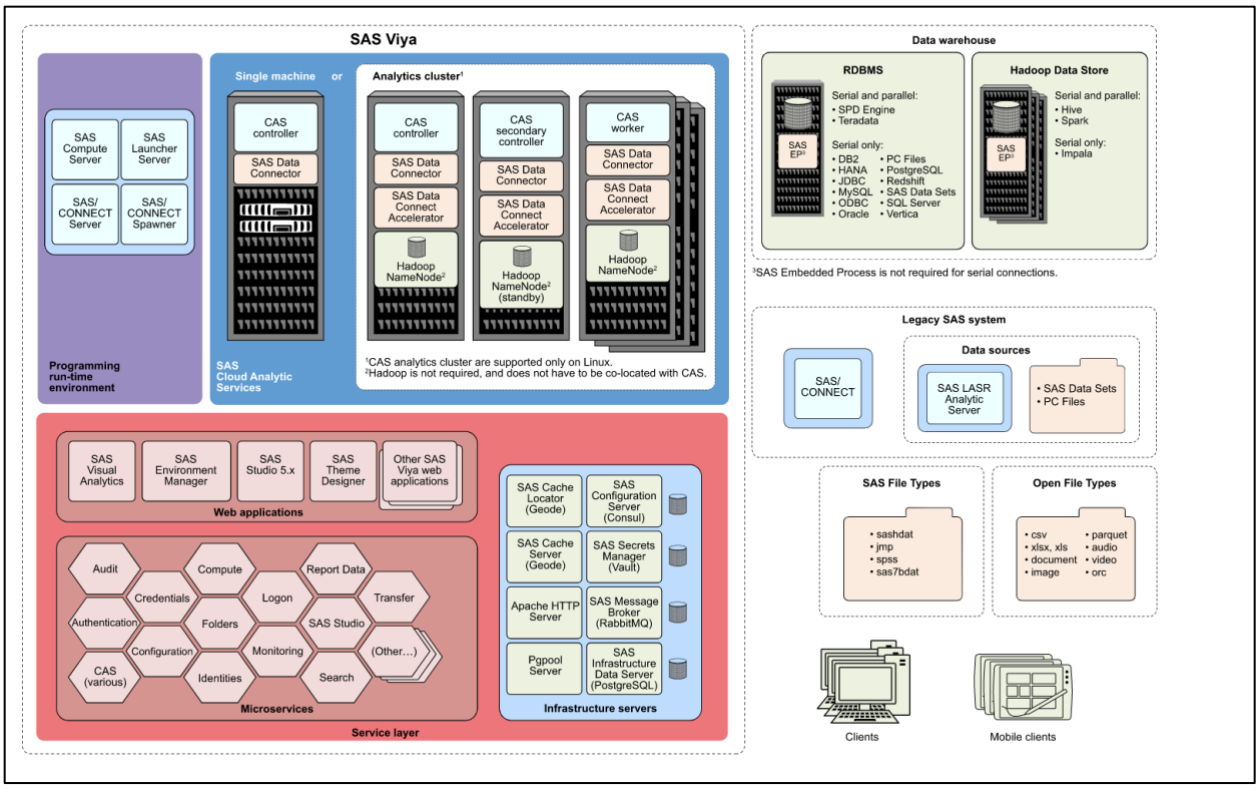

Figure 3: SAS Viya Architecture

Prerequisites for Migrating SAS to AWS

Customers must have a solid understanding of SAS workload requirements and the hardware infrastructure required to meet service objectives – specifically time to complete the task. Existing SAS customers can use the following prompts to assess their understanding:

- Are there any SAS jobs that must run within a certain timeframe? Do you expect your SAS jobs to execute in the same amount of time (or faster) than they are currently executing in your existing data center? If you do expect a similar execution, you should determine the AWS I/O throughput.
- What is the location of the source data for the SAS job? If the data is not in AWS, you must consider the time connection requirements for migration. Added time will impact the SLA for jobs that consume data outside of AWS.
- Is additional security required for the data and/or SAS code?

SAS 9 workloads require instances that supports heavy analytical processing and large sequential I/O; SAS Viya processes data in memory with spill to disk, if required. These

behaviors should be taken into consideration when selecting the correct Amazon EC2 instance types and storage requirements for AWS migration.

Running SAS workloads on the cheapest AWS EC2 instances does not necessarily provide the best performance. For example, customers may require storage and server instances with more physical cores than required for computing needs and more storage capacity than the initial sizes required to acquire the maximum I/O bandwidth for their SAS application(s).

Evaluate the following areas to understand the main considerations for optimal AWS performance:

- Instance types for SAS 9.4, SAS Viya, and/or Hybrid
- Ephemeral, persistent, and shared storage types
- Shared file system for SAS Grid Manager
- Placement of SASWORK, CAS_DISK_CACHE, and permanent SASDATA
- High Availability, Security, and Authentication requirements

Instance Types

Physical Core Requirements

SAS 9 and SAS Viya infrastructures require several SAS server types, depending on usage patterns. Each infrastructure has different and specific CPU, I/O throughput, and memory provisioning. Each instance contains a default number of virtual CUPs (vCPUs) that include CPU core and hyper threads for each core, which allows for multiple-threads to run concurrently on a single CPU core. The number of vCPUs helps determine the total number of physical cores required for SAS.

Additional Configurations

SAS servers require high I/O and sufficient bandwidth is required. You can obtain bandwidth through a dedicated network interface card (NIC) — additional CPUs and RAM may also be required.

Multi-tenancy

With a dedicated NIC, sharing the same with multi-tenant applications residing the same physical server can result in inferior performance for virtualized EC2 instances.

Placement Groups

Ensure that all the instances and components of SAS are placed on the target infrastructure within the same Availability Zone (AZ) in an EC2 placement group. This is particularly useful for SAS workloads that require low-latency performance for node-to-node communication.

SAS 9 Systems

SAS Compute Tiers + SAS Grid Node

SAS compute and SAS grid nodes require a minimum of 8GM of physical RAM per core and robust I/O throughput.

SAS WORK is a temporary library that is automatically defined by SAS at the beginning of each SAS session or job. The WORK library stores temporary SAS files that are created by users and is internally used by SAS. SAS UTILLOC is a temporary location for operations such as sorting, stats, multi-threaded processes which could have the same location as the WORK folder, but may be different.

The following servers are recommended:

- **I3 family** – EC2 I3 instances are the next generation of storage optimized instances for high transaction, low latency workloads. These instances include Non-Volatile Memory Express (NVMe) SSD-based instances storage optimized for high random I/O performance, high sequential read throughput, and high IOPS. Because of the high internal I/O bandwidth from striped NVMe SSD drives for SAS WORK and SAS UTILLOC, users should configure their environment to explicitly use the NVMe SSD local drives (not EBS volumes).
- **I3en family** – This family provides Non-Volatile Memory Express (NVMe) SSD instance storage optimized on Amazon EC2 with enhanced networking via ENA to achieve up to 100 Gbps of network bandwidth.
- **M5n family** – The M5 family provides a balance of compute, memory and networking. M5n instance variation are ideal for applications requiring improved network throughput and packet rate performance.

Shared File System Storage Required for SAS Grid

These servers need robust I/O throughput to the permanent storage that supports the shared file system. They need 8 GM of RAM per physical core.

- **M5 or M5dn family** — With M5dn instances that support 8 GB of RAM per physical core, local NVMe-based SSDs are physically connected to the host server and provide block-level storage for lifetime of the instance.

Both instances types are suitable for workloads requiring a balance of compute, memory and networking resources.

SAS Mid-Tier and Metadata Servers

These servers do not require compute-intensive resources or robust I/O bandwidth, but they do require more memory than the SAS computing tiers. The recommendation is 24 GB of physical RAM, or 8 GB of physical Ram per physical core (whichever is larger).

- **R5 or R5d family** – R5/R5d instances are suitable for memory intensive applications such as in-memory caches, mid-size in-memory databases and real-time big data analytics.

SAS Viya Servers

Typically, SAS Viya deployments contain the following:

- **Cloud Analytics Services (CAS)** — an analytics engine
- **CAS Server monitor** — web application for basic administration
- **SAS Environment Manager** — web application for enterprise administration
- **SAS Studio** — web application for writing code
- **SAS Visual Analytics** — web application for visual reporting, exploration, and modeling

Some SAS Viya configurations (single, non-distributed node going against small data from a source other than Hadoop) will not need all the servers listed here. Additional workload assessments may be necessary to support SAS Viya in AWS.

SAS CAS Nodes – Minimum of Three

SAS CAS are the in-memory analytics engines that support the demands of powerful analytics that require large amounts of data loaded in CAS memory space. Analytical

capabilities can be scaled to add more host machines to the cluster, which allows large processes/analytical problems to be broken down into chunks that are processed by CAS simultaneously across machines.

CAS has the ability to protect against the unexpected loss of worker nodes through a CAS Disk Cache. This cache is a dedicated space on disk that CAS can use as a temporary backing store for data in-memory. If a CAS worker terminates unexpectedly, the other workers can recover the lost allocation of data from the block copies in their own cache on disk.

These servers require fast CPUs for processing data, enough physical RAM to hold all the data files to be analyzed by all the concurrent SAS Viya users, and robust I/O throughput (especially to CAS_DISK_CACHE). If you are not sure how much data will be accessed at any given time, but you know your SAS users will access files in 100s of GBs in size, we recommend 64 GB of RAM per physical core. The following are the recommended instance types for CAS nodes:

- **I3 family** – EC2 I3 instances are the next generation of storage optimized instances for high-transaction, low-latency workloads to support very fast access to CAS_Disk_Cache. These instances include Non-Volatile Memory Express (NVMe), SSD-based instances storage optimized for high random I/O performance, and high sequential read throughput and high IOPS. The high internal I/O bandwidth from striped NVMe SSD drives for SAS Work and SAS UTILLOC is critical.
- **I3en family** – This series provides Non-Volatile Memory Express (NVMe) SSD instance storage optimized on Amazon EC2 with Enhanced networking through ENA to achieve up to 100 Gbps of network bandwidth. Customer can choose this family for higher ephemeral storage requirements
- **R5 or R5ad or R5n family** — These instance types deliver additional memory per virtual CPU. With the R5d instance, local Non-Volatile Memory express SSDs are physically connected to the host server and provide block-level storage that is coupled to the lifetime of the R5 instance.

MicroServices Node

In SAS Viya, applications such as SAS Environment Manager, SAS Visual Analytics, and SAS Logon use microservices for specific, well-defined functionality, which helps the application withstand failures and capacity spikes. All microservices communicate with each other using industry standard conventions, such as REST or HTTP,

regardless of whether the client is a web application or a batch script. All clients use the same mechanism.

Scaling is achieved by allowing individual microservices to scale independently by deploying additional instances and registering them with the SAS Configuration Server. It is best practice to have instances distributed across multiple machines.

Microservices nodes do not require high computational speed or power. To run all SAS Visual Analytics products with SAS Viya, you'll need 96 GB of RAM per node and the following AWS instances types:

- **R5 or R5ad or R5n family** — These instance types deliver additional memory per virtual CPU. With the R5d instance, local Non-Volatile Memory express SSDs are physically connected to the host server and provide block-level storage that is coupled to the lifetime of the R5 instance.

SAS Programming Run-Time Node

This server runs the SAS 9 code base on the application and needs fast CPUs for processing with at least 16 GB of RAM per physical core and robust I/O throughput for SAS WORK and SAS UTILLOC.

- **I3 or I3en family** – As mentioned above I3 instances are best suited for this workload. I3en instances are suitable for workloads with NVMe storage needs above 1 TB upto 60 TB.

Storage Types

AWS has many storage types for temporary and permanent requirements. In this section, we address the options for field experiences and lab testing with the SAS on AWS storage options.

Permanent SAS Data Storage

Permanent SAS storage is used for SAS 9.4, SAS data files, and SAS Viya CAS tables. SAS 9.4 data files hold either a SAS dataset holding actual data, or a SAS non-materialized view definition that references data stored elsewhere.

Viya CAS is not only an analytic and transformation engine, it is also a data server. It loads data into a CAS table in order to analyze and process. The format of these tables

can vary including SASHDAT, CSV, Oracle, SQL Server, or Hadoop, and it is backed into a permanent storage with content stored in-memory.

The following permanent storage options are suggested to support the SAS data files and SAS Viya CAS tables:

- **Elastic Block Storage (EBS)** – Stripe together a minimum of 4 EBS volumes for I/O bandwidth aggregation.
    - EBS ST1 (throughput optimized HDD) Storage designed for large block sequential I/O. A 12.5 TB volume can sustain 500 MB/second. If the volume size is less than 500 MB per second of total bandwidth, it can be observed during the burst window.
    - For high-throughput read-heavy workloads (like in SAS), update the read-ahead setting on EBS ST1/IO1 from default 256 KB to 8 MB. EBS IO1 (provisioned IOS SSD) storage can also be used.
    - Other EBS storage types like GP2 (general purpose storage) and SC1 (cold storage) are not suitable for permanent SAS 9 or SAS CAS data files.
    - RAID 0 configuration is preferential because fault tolerance is not a determining criterion for these workloads.
    - Customers can also choose to have EBS IO1 volumes (provisioned storage). However, costs would increase as IO1 volumes are charged by storage and by provisioned IOPS. For ex – 32K IOPS can yield as much as 500 MB/sec but customers would pay an additional amount for the desired provisioned IOPS.
- **S3**
    - Using SAS/Access to Redshift, SAS Datasets can be loaded into S3/Redshift using AWS S3 capabilities of multi-part upload, transfer acceleration, and COPY/UNLOAD to Redshift for relational storage.

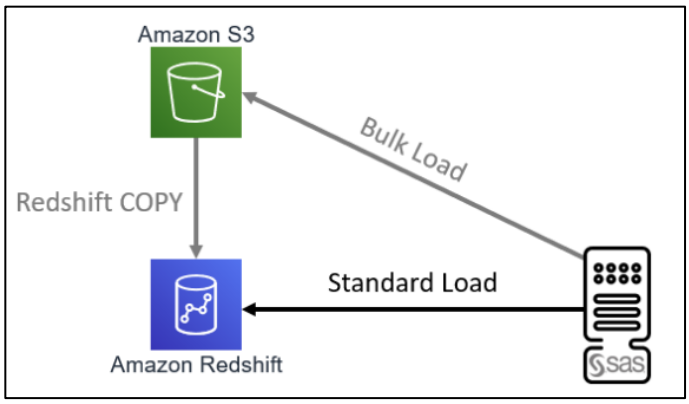

Figure 4: SAS – Redshift Bulk Load

Temporary SAS Data Storage

Most temporary SAS storages used in SAS WORK, SAS UTILLOC, and CAS_DISK_CACHE will not persist through reboots and are considered ephemeral storage.

- I3 instances feature low latency NVMe SSDs striped together with RAID0. Use NVMe devices to support high bandwidth, low latency, and sequential I/O .
- If additional storage is required, default to permanent SAS storage.

Shared File System to use SAS Grid Manager

SAS Grid Manager requires a shared file system for the permanent files being shared by all the SAS Grid compute nodes. The following options have been tested with SAS Grid manager on AWS:

- **FSx for Lustre** — Provides a high-performance file system optimized for fast processing of workloads such as machine learning and high performance computing. These workloads commonly require data to be presented through a fast and scalable file system interface and typically have data sets stored on long-term data stores like S3.

  AWS sets up the Lustre file system with mount options *noatime* and *flock*. SAS prefers flock parameter and the mount options must be properly implemented for FSx for Lustre.

FSx does not allow dynamic expansion for the size of the Lustre File System. If a larger size is required then a new system must to be setup, and data must be copied to the new file system.

- **Elastic File System (EFS) —** EFS supports network file system version 4 protocol and allows multiple EC2 instances to interact with EFS. However, the maximum throughput I/O is 250 MB per second per instance. For more information on limits, refer to the EFS documentation.

  Multiple EFS file systems per instance are required to overcome this I/O throughput limitation in addition to a single NIC per AWS EC2 instance. These file systems cannot be striped together. These file systems have a total of 512 hard locks for any particular file across all users and instances connected to this system.

Placement Groups in SAS

Cluster placement groups are recommended for all SAS core infrastructure components that benefit from low network latency, high network throughput, or both, and if the majority of the network traffic is between the instances in the group. To provide the lowest latency and the highest packet-per-second network performance for your placement group, choose an instance type that supports enhanced networking such as I3 instances. While it is possible that there could be some degradation of performance, partition placement groups and spread placement groups help reduce impact from the likelihood of correlated hardware failures for the application.

In a partition placement group, each group is split into logical segments called partitions containing their own set of racks with separate power supplies and networks. This creates a hardware resiliency in case of failure. A partition placement group can have partitions in multiple Availability Zones in the same Region.

Clients

The most common placement of SAS clients like SAS Enterprise Guide, SAS Data Integration Studio and SAS Studio should be within the same Region where the other SAS infrastructure is located. These clients can be on a Windows server or a Windows Virtual Desktop, and it is required to determine a place for these Windows systems.

Depending on the volume of data being transferred back to the SAS client, having the clients and backend server in the same Availability Zone, Placement Group or Region yields the fastest results. SAS clients can be placed on the same instance (Windows

Server) as the backend server as well, allowing SAS users to access both the clients and the backend server.

SAS Infrastructure

SAS has several tools that allow sharing of SAS data files on-premises with SAS applications that run on AWS and vice-versa. Network bandwidth over the internet can be limited, sometimes as low as 500 KB/second, which can further constraint the I/O required by SAS applications

If higher I/Os are desired, you can use AWS Direct Connect, which is a private network connection between AWS and your datacenter or corporate network that provides a high bandwidth, consistent network experience. It is best practice to have SAS applications and the frequently accessed data in the same Region and Availability Zone.

Source Data Files

It is best practice to co-locate the source data files in the same Region and Availability Zone as the systems using those files. Many SAS GUI based clients pull data from the source data files when populating screens, so having the source data files across slow network connection to AWS will greatly impact performance of the SAS client.

Authentication Tools

If the interaction is frequent, it is recommended to co-locate the authentication tool in the same Region/Availability Zone/placement group. However, in many cases once access is obtained interaction is infrequent with authentication tool and the tool can be located either on-prem or in another availability zone.

Options for Deploying on SAS on AWS

Quick Starts are built by solutions architects and partners to help customers deploy popular technologies on AWS, based on AWS best practices for security and high availability. These accelerators reduce hundreds of manual procedures into just a few steps, so customers can build production ready environments quickly and start using them immediately.

Quick Starts for SAS Grid

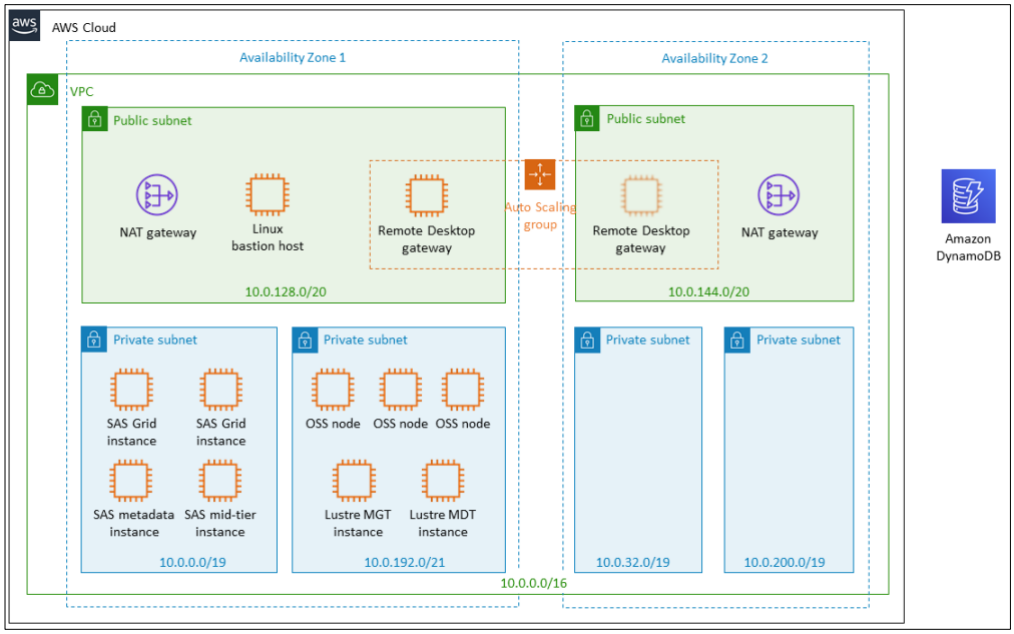

Figure 5: SAS 9 Grid Quick Start

[Quick Starts for SAS Viya]

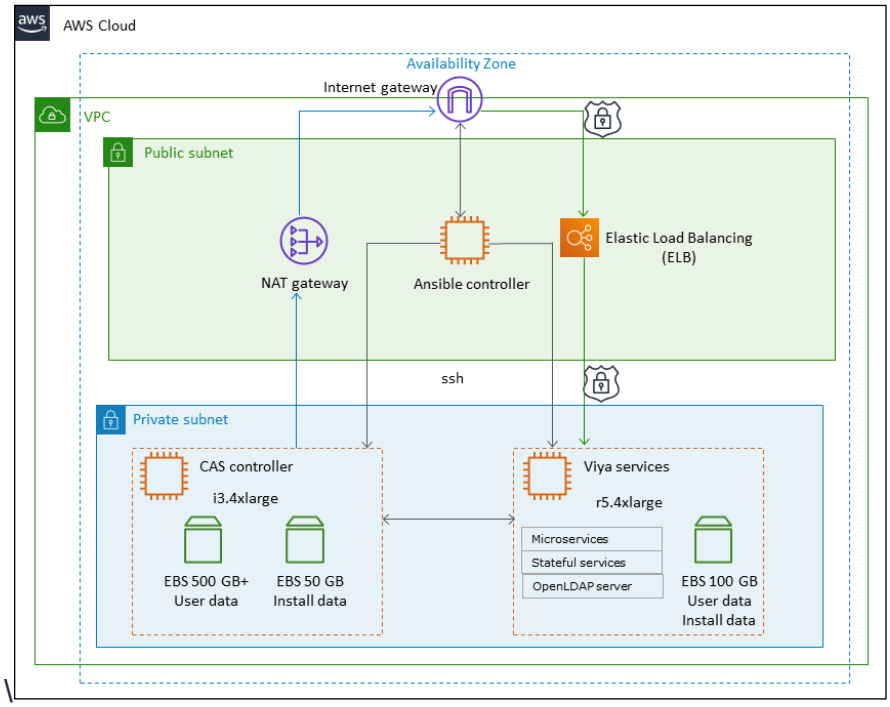

Figure 6: SAS 3.4 Viya Quick Start

Authentication

For Windows server, AWS provides AWS Directory Service or AD connector for connecting an existing on-premises Microsoft Active Directory. For added security MFA can be enabled through Enable MFA on AD Connector.

High Availability

Depending on the need for high availability (HA), a process should be placed in a new instance in case that the existing instance fails. This failover practice is a benefit of running SAS application on AWS

For SAS Grid Manager customers, a shared file system, such as Intel Lustre or IBM Spectrum Scale, will remain operational if one of the nodes associated with the shared file system terminates unexpectedly. However, any data associated with the node will not be available until the node is restored. Only one copy of the data is stored by default and it is possible to enable replication of the data to two or three copied across the

shared file system from the nodes in the system, especially with FSx Lustre, which is backed up into S3.

If customers would like to implement redundancy mechanisms, they must decide the downtime SLAs that they are willing to accept, and based on those choices, implement a pilot light, cold start, warm-standby, or active-active setup. With any of the above HA options, customers must mirror SAS deployment, SAS Files, and data store to the appropriate Region/Availability Zone for HA. At minimum, it is expected that customer builds for cold starts with data stores and deployment files are backed up to S3 in separate production accounts. For more information, refer to the option for single host SAS 9.4 on AWS with backups to Amazon S3 Glacier.

Network and Security

This section covers network and security considerations for SAS deployment on AWS.

SAS 9.4 can be deployed within a customer's VPC within a private subnet containing the required EC2 instances and permanent storages devices.

A public subnet can contain a NAT Gateway that allows instances in a private subnet to connect to the internet or other AWS services, while also preventing outside internet connection to the SAS server.

A bastion host can be placed within the public subnet, with security group rules, to allow transfix between the public bastion host and the SAS servers placed in the private subnet.

Internet Gateway can be used for connectivity between the internet and SAS Servers in a VPC for hosting public websites

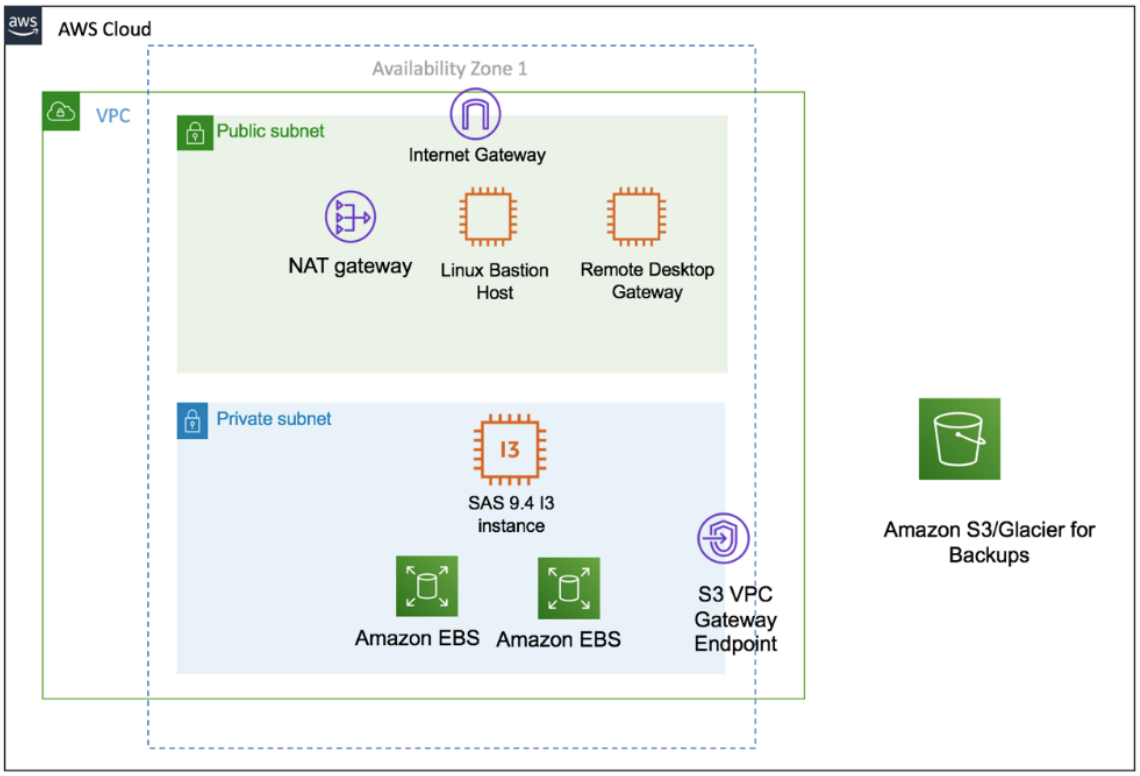

Figure 7: SAS 9.4 Intelligence Platform on AWS

Conclusion

This whitepaper helps SAS customers and migration consultants understand the different AWS capabilities that best meet the needs of their SAS workloads. They key takeaway is that I/O throughput is crucial and can be a limiting factor in successful SAS deployments on AWS if not implemented correctly.

At the same time, it is important for AWS migration consultants and solutions architects to broadly understand SAS infrastructure and its components, so that they can match the base performance requirements with the corresponding infrastructure and technical improvements being released by AWS.

This whitepaper raises awareness of SAS workload requirements within its core and ancillary components, and how to best meet those requirements in AWS. It is crucial to understand that the choice of compute, storage, and application architecture placement are key for achieving the best performance that AWS can offer for SAS deployments.

The information in this paper is based on customer experience and expertise from SAS and AWS working together at the time of writing of this paper. AWS offerings are constantly improving, and therefore it is in the best interest of the reader to understand the rationale used in the selection process and to consider what was done as a point-in-time design.

Contributors

Contributors to this document include:

- Margaret Crevar, Sr Manager, SAS Performance Labs
- Dilip Rajan, Partner Solutions Architect, Amazon Web Services
- Francesco Marelli, Sr Solutions Architect, Amazon Web Services
- Sathish Jothikumar, Sr Product Manager, Amazon Web Services

References & Further Reading

- Important performance Consideration for SAS on Public Cloud
- SAS 9.4 Intelligence Platform
- SAS 9.4 Grid
- SAS Viya
- Five approaches to High-performance Data Loading to the SAS CAS Server

Document Revisions

Date	Description
Feb 2020	First publication