TASI

TELECOMMUNICATIONS AND SOCIAL INFORMATICS
RESEARCH PROGRAM

Massively Learning Activities

Full Name	Student ID
In Woo Park	25141090

May 18, 2023

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Massively Learning Activities (MLA) is divided into two phases, (1) the initial deployment of SAS on existing infrastructure and later (2) the migration of SAS onto scaled infrastructure.

Table of Contents

1 | Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1.1 | TASI/PHIDC

ABOUT US
The Telecommunications and Social Informatics Research Program / Pacific Health Informatics and Data Center (TASI/PHIDC), formerly TASI/PEACESAT, is part of the Social Science Research Institute (SSRI) of the College of Social Sciences (CSS) at the University of Hawai'i at Manoa. TASI/PHIDC programs incorporate an interdisciplinary approach to education and research, and work with partners from across the University of Hawai'i system, State of Hawai'i and other government and academic institutions from the Asia and Pacific Islands region. Program and research focus areas include policy, planning, information and communications technologies and systems, health information technology, health informatics in Hawai'i and the Pacific Islands region.

MISSION
The TASI/PHIDC Research Program missions are to: (1) Provide technical assistance in policy, program planning and evaluation; (2) Facilitate public and private sector collaboration to improve community resiliency, sustainability, and health system performance; and (3) Build capacity in information technology, health data management, analytics, and data sciences.

FACULTY RESEARCH
TASI/PHIDC conducts interdisciplinary and applied research and provides policy, program, technical assistance, education, and training in Hawai'i and the Pacific Islands Region related to:

- Accessible and affordable Information and Communication Technology (ICT)
- Health Information Technology (HIT)
- Electronic Health Record (EHR)
- Healthcare and claims data management, analytics, and programs
- Telehealth
- Meteorological and disaster communications

1.2 | TASI & CNMI (Contract Explained)

TASI/PHIDC is a Technical Assistance and Research Partner or "TARP" who has an Intergovernmental Cooperative Agreement (ICA) with the Commonwealth of the Northern Mariana Islands (CNMI) State Medicaid Agency (SMA) to design an infrastructure that would allow advanced data analytics and parallel processing of Protected Health Information. After careful consideration, TASI/PHIDC has opted for SAS technologies in a hyper-converged infrastructure.

- Modernize data archive and storage (paper to electronic) of PHI data.
- Want to perform data analytics and machine learning.
- Used RCUH funds to purchase SAS license.
- Therefore, SAS needs to be accessible to multi-tenants and UH themselves.

1.3 | TASI & SAS (Contract Summarized)

1. Pre-Deployment and Project Management (ETC 14 Hours)

   - Before deploying SAS technologies, TASI and SAS will engage in pre-deployment and project management tasks.

   - These tasks will involve ongoing project management to ensure that the project plan is followed, and appropriate resources are assigned. The project plan will include details of billable work hours logs that will be sent by SAS and verified by UHTASI. In addition, SAS will send Pre-Install Requirements Documents to UHTASI for completion, and UHTASI will review the completion of these documents to ensure environmental readiness for installation. These tasks will require an estimated 14 hours of work.

2. Deployment (ETC 70 Hours)

   - During the deployment phase, TASI will receive the installation of several SAS products:

     - SAS Advanced Analytics for Education (on Viya 3.5)
     - SAS Data Preparation
     - SAS Data Management Advanced
     - SAS Education Analytical Suite
     - SAS Text Analytics for Education

   - Configuration will also be performed, which includes establishing a database connection and testing it. A validation of the new environment will be conducted to ensure that all components are working as intended before the handoff. Data libraries will be created, and SAS user access controls will be established. TASI will also verify that each of the server components is active and is handling requests. Finally, SAS will provide TASI with installation documentation.

2 | VMware

VMware is a company that specializes in developing technologies for virtualization and cloud computing. Its software products and services enable organizations to efficiently manage their IT infrastructure, improve performance, and reduce costs. VMware offers solutions for network virtualization, cloud management, digital workspace solutions, and security solutions.

2.1 | vSphere 6.5

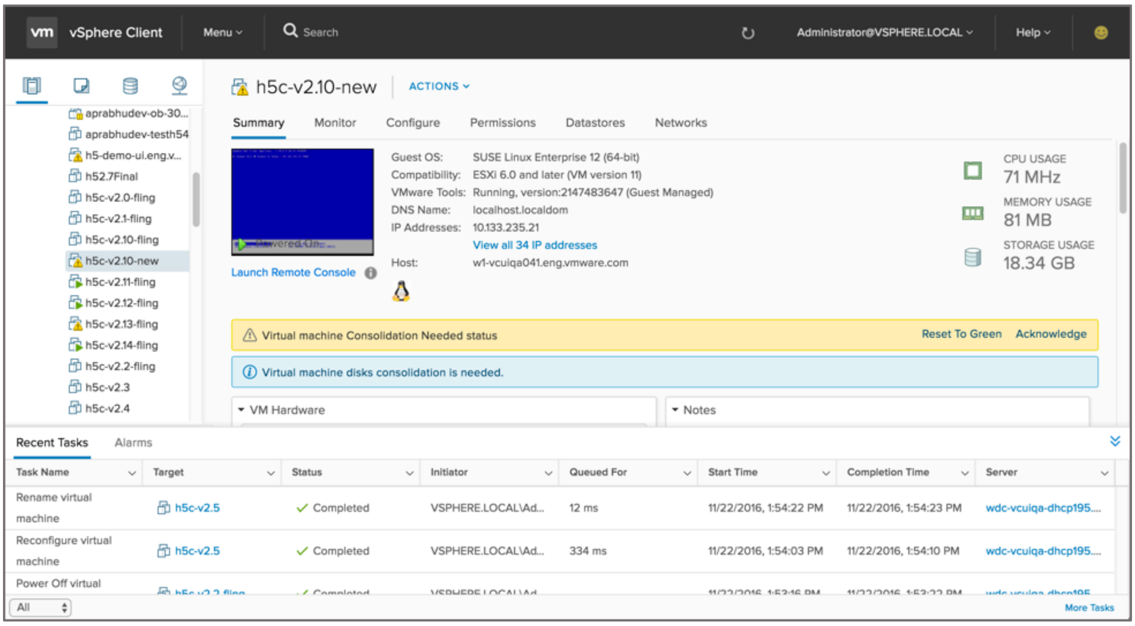
vSphere is VMware's virtualization software suite that allows you to create and manage virtual machines and computing environments, using a set of software tools and services. With vSphere, you can run multiple virtual machines on the same physical server, each running its own operating system and applications. vSphere includes many features and capabilities that help make virtualized environments more reliable, scalable, and performant, such as:

- **vSphere Web Client**: A web-based management interface.
- **ESXi**: The bare metal hypervisor installed on your machines.
- **vCenter**: A centralized management system for your vSphere environment.
- **vSAN**: A software-defined storage solution to create a distributed storage platform in vSphere.
- **NSX**: A software-defined networking solution for your vSphere environment.
- **VMotion:** Software to migrate VMs between servers without interruption of service.

2.2 | vSphere Web Client

The vSphere Client is an application that enables administrators to manage and monitor VMware vSphere environments. It comes with a graphical user interface (GUI) and allows users to connect to VMware vCenter Server, which serves as a central management console for multiple VMware vSphere hosts.

Through the vSphere Client, administrators can create and modify virtual machines, manage storage, configure networking, and monitor system performance, among other things. Essentially, it provides a range of tools that enable users to manage virtual infrastructure components effectively. In addition to the traditional Windows-based vSphere Client, there's also a web-based version called the vSphere Client (HTML5), which is designed to work seamlessly across different operating systems and devices, including desktops, laptops, and mobile devices. This new client offers a simplified interface, improved performance, and support for new features introduced in vSphere 6.5 and later versions.



h5c-v2.10-new

Guest OS: SUSE Linux Enterprise 12 (64-bit)
Compatibility: ESXi 6.0 and later (VM version 11)
VMware Tools: Running, version:2147483647 (Guest Managed)
DNS Name: localhost.localdom
IP Addresses: 10.133.235.21
Host: w1-vcuiqa041.eng.vmware.com

CPU USAGE 71 MHz
MEMORY USAGE 81 MB
STORAGE USAGE 18.34 GB

Virtual machine Consolidation Needed status

Virtual machine disks consolidation is needed.

Task Name	Target	Status	Initiator	Queued For	Start Time	Completion Time	Server
Rename virtual machine	h5c-v2.5	✓ Completed	VSPHERE.LOCAL\Ad...	12 ms	11/22/2016, 1:54:22 PM	11/22/2016, 1:54:23 PM	wdc-vcuiqa-dhcp195....
Reconfigure virtual machine	h5c-v2.5	✓ Completed	VSPHERE.LOCAL\Ad...	334 ms	11/22/2016, 1:54:03 PM	11/22/2016, 1:54:10 PM	wdc-vcuiqa-dhcp195....
Power Off virtual	h5c-v2.2-fling	✓ Completed	VSPHERE.LOCAL\Ad...		11/22/2016, 1:53:16 PM	11/22/2016, 1:53:22 PM	wdc-vcuiqa-dhcp195

Figure 2.1: vSphere Client (STOLEN EXAMPLE)

2.3 | ESXi

VMware ESXi formerly known as ESX is a bare metal hypervisor that is installed directly on the physical server hardware and provides the ability to create, run, and manage virtual machines.

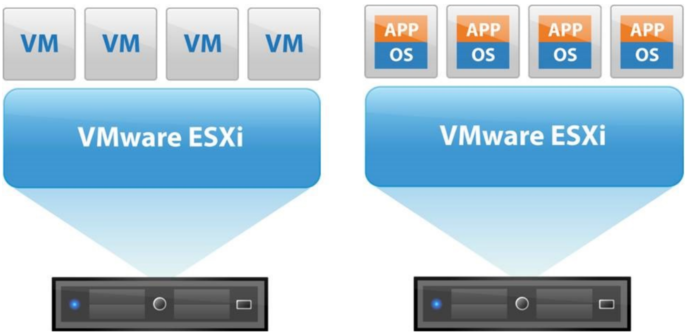


Figure 2.2: ESXi (STOLEN EXAMPLE)

2.4 | vCenter

vCenter is a software platform that provides centralized management and control for their suite of virtualization products, including vSphere. By providing a single point of control, it simplifies management and reduces complexity, making it easier to manage many virtual machines and components. With vCenter, you can manage hosts, clusters, virtual machines, networks, and storage resources to support a virtualized environment with high availability, disaster recovery, and workload balancing.

In addition, vCenter provides advanced capabilities like automation, orchestration, and policy-based management. These features allow you to automate routine tasks, streamline operations, and enforce policies across your virtualized environment. Examples of automated tasks include: provisioning VMs [1], patches and updates [2], backup and recovery [3], monitor and reports [4], and resource allocation [5].

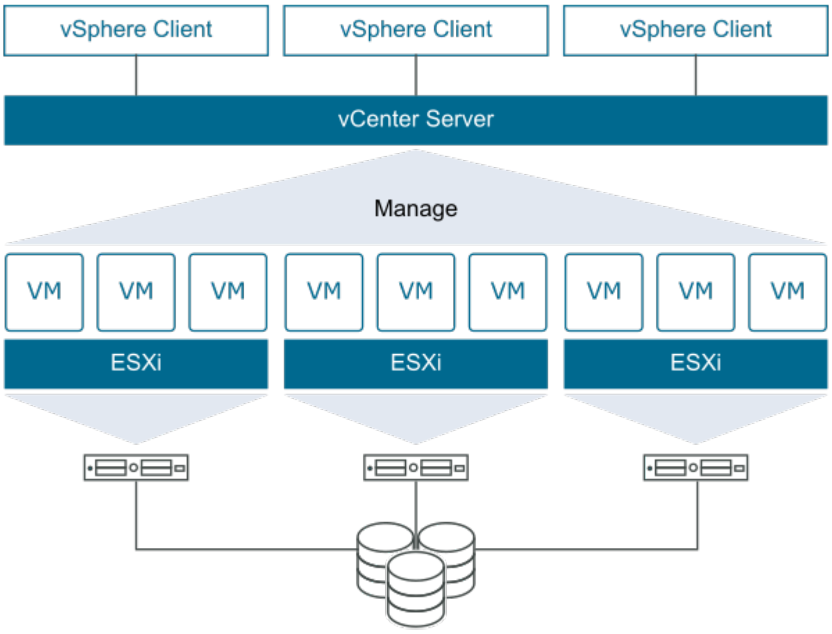


Figure 2.3: vCenter (STOLEN EXAMPLE)

[1] Create new virtual machines, configure virtual hardware, and install operating systems and applications.
[2] Deploy software updates, apply security patches, and performing maintenance tasks.
[3] Create backup schedules, perform backup and restore operations, and monitor backup performance.
[4] Generate reports on virtual machine performance, track resource usage, and monitor system health.
[5] Adjust CPU and memory resources, configure storage allocations, and manage network bandwidth.

2.4.1 | vCenter Security and Risks

Security is a critical aspect of virtualized environments, and vCenter provides a range of security features to protect against unauthorized access, data theft, and data manipulation. These security features include: role-based access control[6], auditing[7], encryption[8], secure communication[9], integration[10], and two-factor authentication[11]. These security features help to ensure confidentiality, integrity, and availability of the virtualized infrastructure, a requirement when working with PHI data.

2.5 | vSAN

vSAN is a software-defined storage solution developed by VMware, which allows organizations to create a distributed storage platform that is integrated with vSphere. This provides a highly scalable and available storage infrastructure, using standard hardware.

By creating a shared data store using the internal disks of ESXi hosts in a vSphere cluster, vSAN allows organizations to pool their storage capacity and performance into a single datastore, scaling it easily by adding more hosts to the cluster. vSAN features data replication, erasure coding, and automatic data rebalancing. Additionally, it offers advanced storage services such as deduplication, compression, and encryption, ensuring optimal storage efficiency and security which streamlines storage management, automates routine tasks, and helps to optimize storage utilization and cost savings.

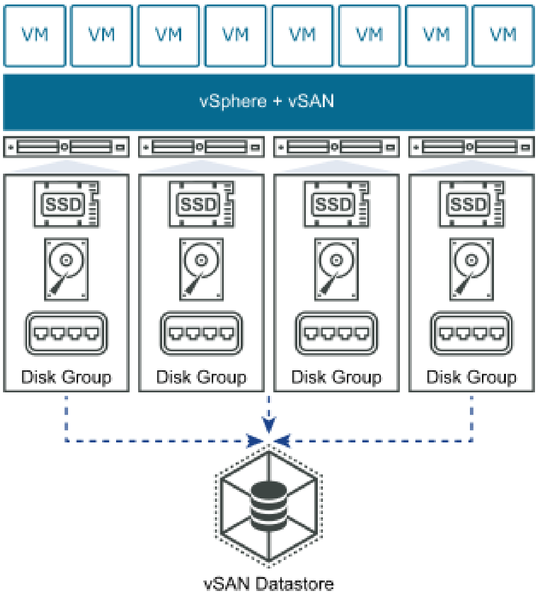


Figure 2.4: Standard vSAN Cluster (STOLEN EXAMPLE)

2.6 | NSX

NSX is a network virtualization and security platform created by VMware that provides a software-defined networking (SDN) solution that enables organizations to virtualize their network infrastructure, creating a more flexible, scalable, and manageable network.

NSX allows for all network components in your infrastructure to be virtualized, decoupling your network from existing hardware. This abstraction enables organizations to pool and automate network resources, which can reduce the time and cost of deploying and managing network infrastructure. NSX also offers advanced security features and networking capabilities which allows administrators to apply precise

[6] Define roles and permissions to users based on their roles to prevent unauthorized access.
[7] Track user activity and changes to identify security issues and log actions taken within the virtualized environment.
[8] Encrypt VM data, configuration files, and communication between hosts.
[9] Supports SSL/TLS encryption to secure communication between hosts and the vCenter server.
[10] Integrate with third-party security products (e.g., antivirus, IDS) to provide additional layers of security
[11] Provide two forms of identification before accessing the VM to prevent unauthorized access.

policies to specific workloads or applications. For example, NSX provides: network automation, multi-cloud and on-premises support, network segmentation, minimal cost and resource overhead, switching and routing, and load balancing features.

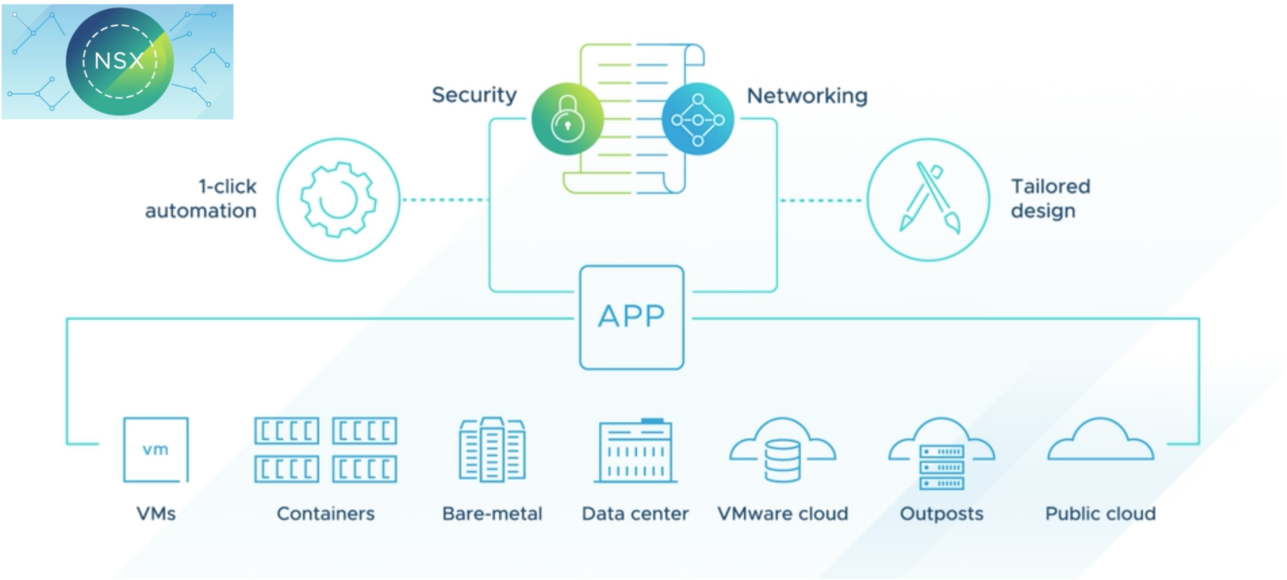


Figure 2.5: NSX Infrastructure (STOLEN EXAMPLE)

2.7 | VMotion

VMotion is virtualization software that enables IT administrators to move VMs between physical servers or hosts without disrupting service. The process involves copying the entire state of the VM, including memory, CPU state, and network connections, from one host to another. The benefits of VMotion include increased availability and uptime, improved hardware utilization, workload balancing, and reduced downtime for maintenance and upgrades. However, the feature also requires specialized hardware and software, increasing the complexity of virtualized environments. VMotion uses shared storage, high-speed networking, and specialized software to ensure a seamless migration.

The main use case for VMotion is to provide high availability and workload balancing for virtualized environments by optimizing resource usage, improving performance, and avoiding downtime during maintenance or upgrades. For example, an IT administrator can use VMotion to move running VMs to another host during server maintenance, ensuring uninterrupted service for end-users. Once the maintenance is complete, the VMs can be moved back to the original host. VMotion also allows for the consolidation of workloads and the migration of VMs to new hosts for improved hardware utilization and cost savings.

When migrating VMs with sensitive data, such as protected health information (PHI), there may be compliance issues with regulations like HIPAA. To ensure compliance, virtualization infrastructure and VMotion must be configured to meet data protection, access control, and auditability requirements. Encryption and other security measures should also be implemented to protect the confidentiality, integrity, and availability of PHI during migration. IT administrators should ensure that host servers and network connections used for VMotion are secure and protected from unauthorized access.

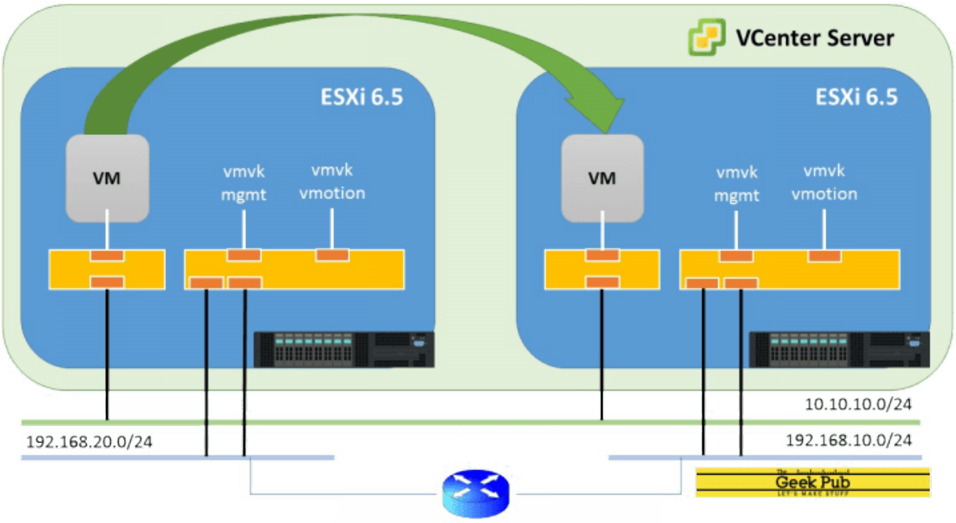


Figure 2.6: VMotion (STOLEN EXAMPLE)

3 | Statistical Analysis System (SAS)

SAS, or Statistical Analysis System, is a software suite that has been used for advanced analytics, business intelligence, data management, and predictive analytics since it was first released in 1976. Developed by the SAS Institute, it offers a range of statistical and data analysis tools, which are suitable for many applications including data mining, forecasting, econometrics, quality control, and statistical analysis.

The software provides a user-friendly graphical interface for data analysis and reporting, as well as a powerful programming language that allows users to customize their analysis and automate repetitive tasks. Its ability to handle large and complex datasets and perform advanced statistical analyses make it popular in various industries, including finance, healthcare, government, and academia. SAS is widely used for purposes such as fraud detection, risk management, clinical research, and marketing analysis, and is a popular choice among data scientists and statisticians.

SAS offers multiple product **suites**. The SAS Enterprise Suite is a collection of SAS products designed for enterprise-level data management and analysis. The SAS Platform provides two engines for managing foundational capabilities such as distributed processing, security, administration, program development and execution, resource management, user interfaces, cloud integration, operating systems and third-party software. These engines are SAS 9.4 and SAS Visual Analytics.

In addition to the SAS Enterprise Suite and the SAS Platform, SAS also offers SAS Data Management Advanced which is a powerful ETL solution for preparing data for both analytic engines.

3.1 | SAS Data Management Advanced (SAS DMA)

SAS Data Management Advanced (SAS DMA), is a software suite that provides a comprehensive set of tools for data integration and data quality. The primary purpose of SAS DMA is to support data ETL (Extract, Transform, Load) processes, which involve extracting data from multiple sources, transforming it to meet specific requirements, and loading it into a target system for analysis and reporting. SAS DMA is a stand-alone software suite and can be deployed on-premises or in the cloud. It is not part of SAS 9.4 (i.e. SAS Data Management and Analytics), which is a separate software suite that provides a wide range of tools for data analysis, reporting, and visualization.

To support these functions, SAS DMA relies on three different types of servers:

- The **Mid-Tier** server is a web-based interface that provides access to SAS DMA workflows. This server is responsible for user authentication and authorization, job scheduling and monitoring, and other functions that are necessary for effective workflow management. It acts as a gateway for users to interact with the SAS DMA system.
- The **Metadata** server is responsible for managing information about data sources and workflows. It provides a central repository for storing metadata, which enables efficient management of SAS objects, definition of relationships between objects, and tracking of changes to data. In the case of SAS DMA, the metadata server manages information about data integration workflows and data quality rules.
- The **Compute** server provides the processing power and resources necessary to run data integration and data quality jobs. This server is responsible for executing the actual data integration and ETL tasks defined in the workflows created in SAS DMA. It ensures that the workflows are run efficiently and effectively, regardless of the size or complexity of the data being processed.

3.2 | SAS 9.4

SAS 9.4 is a software suite that provides tools for data management, statistical analysis, business intelligence, and predictive modeling. SAS 9.4 can handle large datasets and complex analyses by using a wide range of built-in functions and procedures that can save time and effort when working with data.

For example, a pharmaceutical company might use SAS 9.4 to analyze clinical trial data to determine the efficacy and safety of a new drug. A bank might use SAS 9.4 to perform risk analysis on its loan portfolio. A retail company might use SAS 9.4 to analyze customer data to better understand buying patterns and preferences.

SAS 9.4 is composed of several modules that provide a wide range of functionalities:

- Base SAS - The basic programming language, data access, and management capabilities of SAS.
- SAS/STAT - A comprehensive set of statistical analysis procedures for data exploration and modeling.
- SAS/GRAPH - A set of tools for creating high-quality graphical output from SAS data.
- SAS/ETS - A set of time series analysis and forecasting procedures.
- SAS/IML - An interactive matrix language for matrix manipulation, data analysis, and numerical optimization.
- SAS/ACCESS - Connectivity to data sources such as relational databases and spreadsheets.
- SAS Enterprise Guide - A graphical user interface (GUI) for SAS programming, data management, and reporting.

3.3 | SAS Visual Analytics (SAS Viya)

SAS Visual Analytics (SAS Viya), is a cloud-based analytics platform that provides a suite of tools and services for elastic, scalable, and fault-tolerant data analytics, data processing, and machine learning for enterprise environments. It allows organizations to store, manage, analyze, and share large volumes of data across different sources and formats, all within a single platform.

SAS Viya is composed of several software that provide a wide range of functionalities:

- SAS Visual Analytics: A tool for creating interactive reports and dashboards to explore and visualize data.
- SAS Visual Statistics: A tool for performing statistical analysis and building predictive models on large data sets.
- SAS Visual Data Mining and Machine Learning: A tool for exploring and analyzing large data sets using advanced analytics techniques such as clustering, decision trees, and neural networks.
- SAS Visual Forecasting: A tool for creating accurate and reliable forecasts using time series data.
- SAS In-Memory Statistics: A tool for performing high-performance analytics and modeling on large data sets using in-memory processing.

When performing analytics on large datasets, SAS Viya uses Cloud Analytic Services.

3.4 | Cloud Analytics Services (CAS)

Cloud Analytics Services (CAS) is the in-memory analytics engine SAS Viya uses for both on-premise as well as cloud-service environments (e.g., AWS, Azure, GCP). CAS uses a combination of hardware and software where data management and analytics take place on either a single-machine or as a distributed server across multiple machines. In either single or distributed deployment, each machine (host, node, etc) will be assigned one of three roles: CAS Controller, CAS Backup Controller, CAS Worker.

Analogy
In a restaurant kitchen, there exists three primary chefs. They are the (1) executive chef, (2) sous chef, and (3) station chef(s). The executive chef's primary role is to manage the kitchen and its staff whilst doing very little cooking. The sous chef's primary role is to be the right-hand to the executive chef, ready to manage the kitchen, share, or take over the responsibility of the executive chef at a moments notice. The station chef(s) merely wait for instructions from the executive chef, then executes the job they are given.

This is the relationship of each CAS node with each other:

- The CAS Controller is the executive chef managing the kitchen and its staff, delegating work.
- The CAS Backup Controller is the sous chef ready to take over the responsibility of the executive chef.
- The CAS Worker(s) are the station chefs cooking what they are assigned to by the executive chef.

3.4.1 | Role 1: CAS Controller

Controller is the first role that can be assigned to a host for SAS Cloud Analytic Services. For both server architectures, single-machine and distributed, one machine must be designated as the Controller. The role of the Controller is to parse out work to each Worker host available. In other words, the Controller manages and controls the overall operation of the CAS environment. As the master node, the Controller is responsible for distributing workload among available CAS Workers, managing user sessions, and providing a secure environment for data retrieval and data storage.

In a single-machine environment, the CAS Controller and CAS Worker roles can be performed by different processes or threads within the same operating system instance. However, we are not limited to this deployment method as it is also possible to have the CAS Controller and CAS Worker(s) virtually separated (on the same hardware) to increase the scalability of the deployment. The configuration of your architecture depends on what you need out of CAS.

In a distributed environment, the CAS Controller is responsible for managing and controlling the CAS environment whilst the actual data processing and data analytics are performed by the CAS Worker(s).

3.4.2 | Role 2: CAS Backup Controller

Backup Controller is the second role that can be assigned to a host for SAS Cloud Analytic Services. Although optional, the CAS Backup Controller is highly recommended in a distributed server environment. The role of the CAS Backup Controller is to act as a standby or hot-backup for the primary CAS Controller in case of a failure. Its primary purpose is to ensure that the system can continue to function in the event of a failure of the primary controller. The Backup Controller is typically set to passively monitor the primary controller for any signs of failure, such as a loss of connectivity or failure to respond to heartbeat messages. It does not actively participate in task scheduling or job execution while the primary controller is running normally.

If the primary CAS Controller fails, the Backup Controller will take over as the primary controller and assume responsibility for managing the CAS worker nodes and scheduling tasks. In this scenario, the CAS worker nodes will send their status updates and job results to the Backup Controller instead of the failed primary controller.[12]

In some systems, the Backup Controller can also be given jobs to execute as a CAS worker node. This can help to improve the system's overall performance by increasing the number of available processing resources. In this scenario, the Backup Controller can perform both the role of a CAS Controller and a CAS worker node.[13]

3.4.3 | Role 3: CAS Workers

Worker is the third role that can be assigned to a host for SAS Cloud Analytic Services. The CAS Worker is responsible for performing data processes and data analytics sent from the CAS Controller. For example, CAS Workers can perform data manipulations, transformations or computations on large/complex datasets. These computations are but not limited to: statistical analysis, machine learning models, text analysis, time series analysis, optimization, etc. Workers execute these computations using data stored on disk, in-memory, or in a distributed file system.

In a distributed environment, one host will be assigned as your controller and any additional hosts are considered workers (optional CAS Backup Controller). Workers increase the overall computing power of your distributed-server and provides a solution for a scalable (up/down), distributed, and fault-tolerant environment for data storage and data analysis because the worker manages the storage of data/metadata across multiple nodes. The amount of CAS Workers needed to create an optimized distributed environment is highly dependant on data size, computation type, and workload.

We can create two types of CAS configurations: a single-machine environment using symmetric multiprocessing **(SMP)**, or distributed server environment using massively parallel processing **(MPP)**.

[12] If the main CAS Controller fails, how does each CAS Worker respond to the Backup Controller with their completed jobs?

[13] Can the CAS Backup Controller be assigned work as well as passively monitor the main CAS Controller?

3.4.4 | Symmetric Multiprocessing (SMP)

The Symmetric multiprocessing (SMP) architecture is used when you want to run CAS on a single server or virtual machine (VM) with multiple CPU cores. This is called shared-memory architecture because all the CPUs share the same memory. When a job is submitted to CAS in an SMP architecture, it is processed by the worker node(s) in parallel using the shared memory. The results are returned to the controller node, which sends them back to the user.

A typical SMP architecture for CAS might consist of a single VM that serves as both the controller and worker node. The number of VMs required will depend on the size of your data and the processing requirements of your workload. For example, if you have a large dataset or complex analytical workloads, you might deploy CAS on a VM with 8 or 16 CPU cores.

Some examples of use cases for CAS on an SMP architecture include:

- Exploratory data analysis
- Statistical modeling and regression analysis
- Predictive analytics
- Machine learning and deep learning

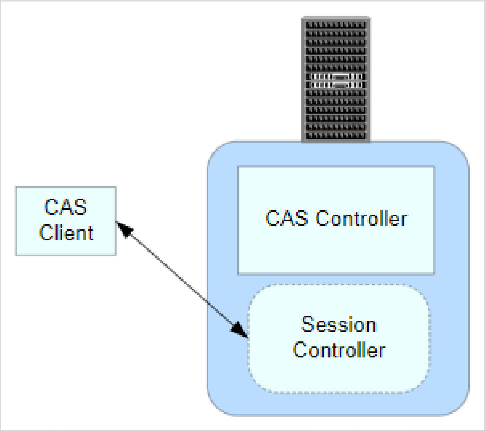


Figure 3.1: Single-machine CAS Server (STOLEN EXAMPLE)

3.4.5 | Massively Parallel Processing (MPP)

The Massively Parallel Processing (MPP) architecture is used when you want to run CAS on a cluster of multiple servers or VMs. This is called distributed-memory architecture because the data is partitioned and stored across multiple servers or nodes. When a job is submitted to CAS in an MPP architecture, it is distributed across the worker nodes in parallel. Each worker node processes its own subset of the data and returns the results to the controller node. The controller node then aggregates the results from all worker nodes and sends them back to the user.

A typical MPP architecture for CAS might consist of multiple VMs or servers, with some dedicated as controller nodes and others as worker nodes. The number of VMs or servers required will depend on the size of your data and the processing requirements of your workload. For example, you might choose to deploy CAS on a cluster of 10 or more VMs or servers to handle large-scale data processing tasks.

Some examples of use cases for CAS on an MPP architecture include:

- Big data processing and analysis
- High-performance computing
- Large-scale machine learning and deep learning
- High-throughput data processing, such as in genomics or drug discovery

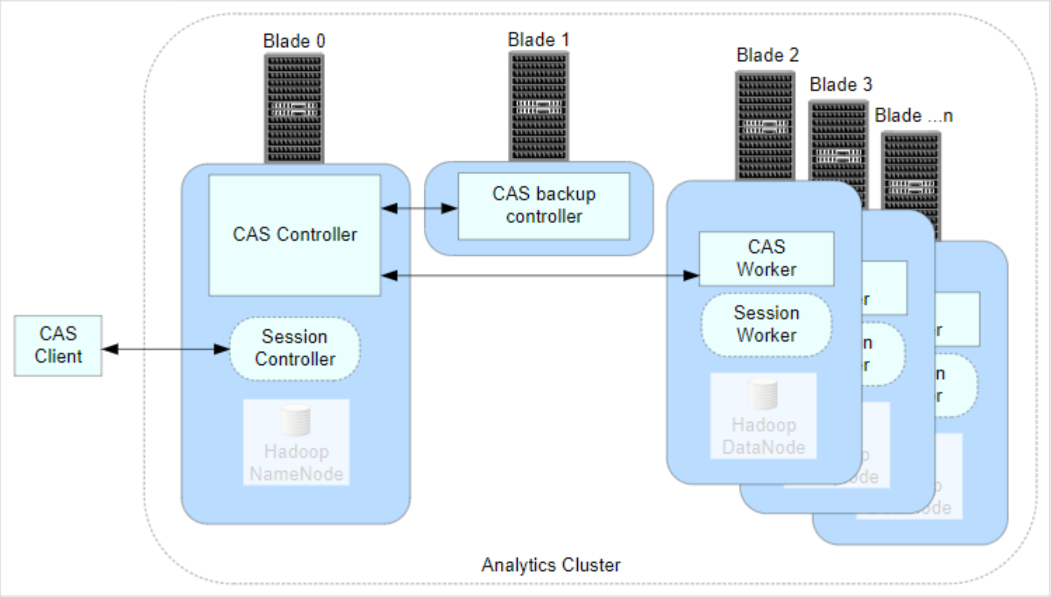


Figure 3.2: Distributed CAS Server (STOLEN EXAMPLE)

4 | Security and Risk Management

This chapter provides an introduction to security and risk management by covering key concepts such as data compliance, identity and access management (IAM), data governance, data encryption, and data backups. This chapter is not intended to be a comprehensive handbook for implementing proper security measures, but rather as an overview of the security measures to consider when developing a strategy for storing and accessing sensitive information.

4.1 | Data Compliance

- HIPAA Compliance
- UHM Compliance
- RCUH Compliance
- TASI Compliance
- State of Hawaii Compliance

4.2 | Identity and Access Management (IAM)

Identity and Access Management (IAM) is a security practice that safeguards sensitive information by allowing only authorized individuals to access confidential resources and data.

Identity management looks to confirm that an accessing user is who they say they are, whilst access management uses a users identity to determine which resource they are allowed to access.

IAM components can be classified into four major categories: authentication, authorisation, user management, and central user repository.

4.2.1 | Authentication

Authentication is a component of IAM in which a user is required to provide sufficient credentials to gain access to an application system.

Sufficient credentials for accessing sensitive healthcare information are defined as authentication methods that comply with the HIPAA Security Rule (Section 5.1). The HIPAA Security Rule requires covered entities to implement multi-factor authentication or an equivalent authentication method for accessing ePHI.

According to HIPAA, the multi-factor authentication method must use two of the following three elements:

- Something you know (Password or PIN)
- Something you have (Smart Card or Security Token)
- Something you are (Fingerprint or Facial Recognition)

Two new additional standards are not required but provide additional authentication methods:

- Somewhere you are (IP Address or Geo-location)
- Something you do (Signature or Gesture)

Once a user is authenticated, a session is created to allow the user to interact with the application system. The session will remain open until the user's task is completed or through termination by other means (e.g., timeout). By centrally maintaining the session of a user, the authentication module can provide single sign-on services.

Single sign-on (SSO) is a mechanism that allows users to authenticate once and access multiple systems or applications without having to re-enter their credentials. SSO simplifies access control and user permissions by providing a centrally managed solution for user authentication policies across all systems. There are several options when deciding on a SSO solution. (e.g., LDAP, OAuth, SAML, RADIUS, PKI, etc).

4.2.2 | Authorization

Authorization is a component of IAM in which a user is given permission to access a particular resource.

This component comes after a user has successfully authenticated to an application system with sufficient credentials. Authorization is performed by checking the resource access request (e.g., web-based application URL), against an IAM policy store and is the core module that implements Role-Based/Attribute-based, access control.

- Role-Based Access Control (RBAC) is a method of access control that assigns roles to users and access permissions to those roles in order to provide a centrally managed solution for authorization.
- Attribute-Based Access Control (ABAC) is a method of access control that assigns permissions based on a user's attributes (e.g., job title, location, department).

The authorization model can provide more complex access control policies other than user role/groups and user attributes (e.g., access channels, time, resource requested, external data, business rules).

4.2.3 | Central User Repository

The Central User Repository (CUR) stores and delivers identity information in order to verify credentials submitted from clients. Identity information is equivalent to user account information (e.g., usernames, passwords, etc). The most common CUR protocol is the Lightweight Directory Access Protocol (LDAP).

LDAP is a protocol for accessing and maintaining distributed directory information services over an Internet Protocol network in order to provide a centrally managed authentication and authorization solution for application systems.

LDAP allows system administrators to manage user accounts, configure access and permissions, and monitor and audit user activity.

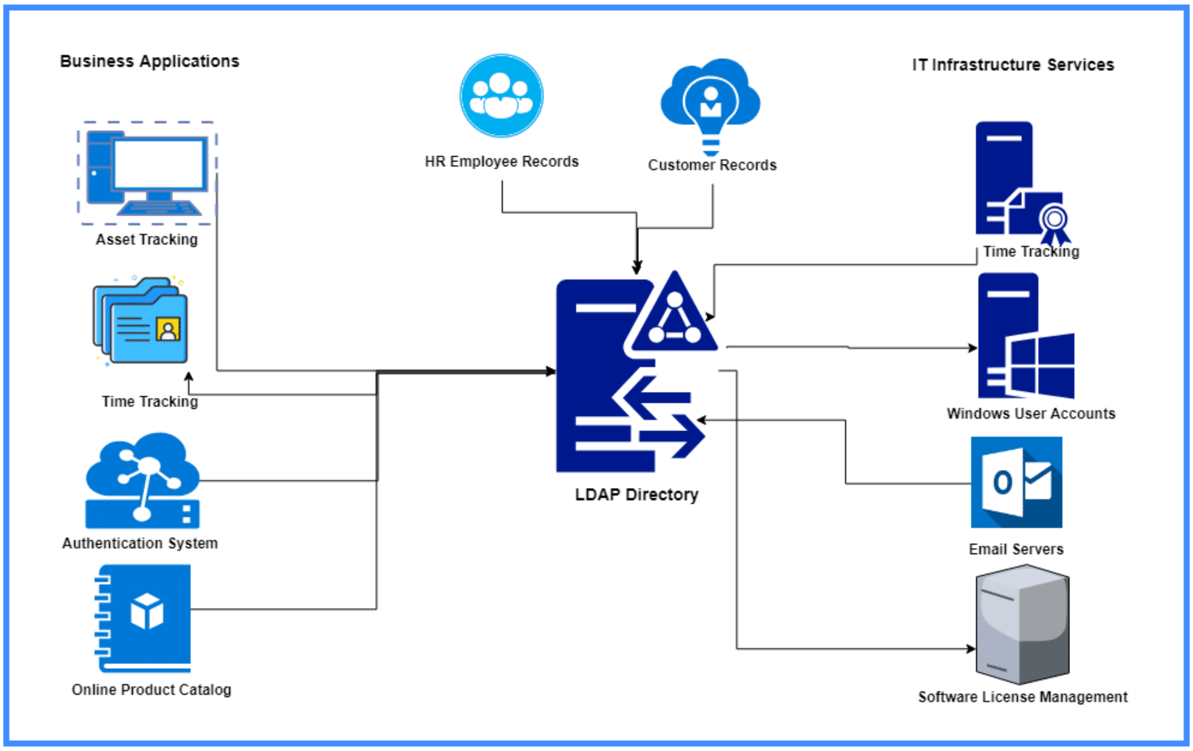


Figure 4.1: Lightweight Directory Access Protocol (STOLEN EXAMPLE)

When designing an LDAP directory, it is important to consider the principle of least privilege. The principle of least privilege is a design principle where each user or service is only given the necessary permissions to perform their intended tasks, and no more. Unauthorized users should not access important data or systems. The principle of least privilege should also be considered when designing security groups and access control lists, as unauthorized users must not have access to sensitive information.

4.2.4 | User Management

User Management is the component of IAM that covers the creation and maintenance of user accounts, account identity, and account privileges.

Identity creation and maintenance is controlled by the set of administrative functions such as user life-cycle management, role/group management, and user/group provisioning. User life-cycle management controls the lifespan of user accounts from account provision to account deprovision. Role/group management and user/group management is used for user authorization (Section 5.2.2).

Onboarding, maintenance, and off-boarding are the three components of user life-cycle management.

1. On-boarding Process:
   - Account authentication for relevant systems and applications.
   - Verification of tenant identity.
   - Setting up multi-factor authentication.
   - Domain and network access configuration.
   - Training for new tenants on how to use the systems and applications they have access to.
2. Maintenance Process:
   - Regular review of tenant access privileges to ensure that they align with the tenant's job function and level of responsibility.
   - Management of access requests and approvals to ensure that access is only granted to authorized tenants.
   - Management of tenant accounts and passwords, including password expiration policies and periodic password resets.
   - Monitoring and auditing tenant activity to detect for potential security threats.
   - Provisioning of additional access or permissions based on changes to the tenant's role or job function.
3. Off-boarding Process:
   - Revocation of tenant access to all systems and applications once life-cycle is expired.
   - Archiving or removal of tenant data in accordance with the organizations (e.g., TASI, RCUH, UHM, etc.) policies and regulatory requirements.
   - Review of tenant access to ensure that no data or resources have been left behind.
   - Disabling or revocation of any credentials associated with the tenant's access.
   - Notification of relevant stakeholders about the tenant's departure.

4.3 | Data Governance

To be completed during the Data Governance Seminar in early May 22-24, 2023.

4.4 | Data Encryption

Data Encryption is a security practice that safeguards sensitive information by transforming the data into an unreadable format that can only be deciphered with the appropriate decryption key.

HIPAA Security Rule (Section 5.1) requires covered entities to implement a mechanism to encrypt and decrypt ePHI based on the assessment of risks to the confidentiality, integrity, and availability of the ePHI.

- Data At-Rest is data that is stored in storage devices (e.g., disk, tap, USB drives, non-votalite storage, etc) and is not being used or transmitted.

- Data In-Transit is data that is transmitted over a network (e.g., file transfers, emails, instant messages). HIPAA requires the use of secure transmission protocols (e.g., SSL, TLS) for transmitting ePHI over public networks.

4.5 | Data Backup

A Data Backup is a copy of data that is used for data restoration in the case of data loss, data corruption, or other data-related disasters.

- Recovery Point Objective (RPO) is the maximum amount of data – as measured by time – that can be lost before data loss exceeds what is acceptable to an organization.
- Recovery Time Objective (RTO) is the maximum tolerable length of time that a system (e.g., can be down after a failure or disaster occurs.

5 | Massively Learning Activities I - Initial Deployment

The System Development Life-cycle (SDLC) is a project management model that defines different stages that are necessary to bring a project from conception to deployment and later maintenance. The SDLC model consists of several phases: planning, requirement gathering, design, implementation, testing & integration, and operations & maintenance. It provides a systematic approach to system development that helps ensure that system is built efficiently with minimal risk.

Massively Learning Activities will follow a similar variation to the SDLC project management model where each SDLC stage will correspond to a subsection in this chapter.

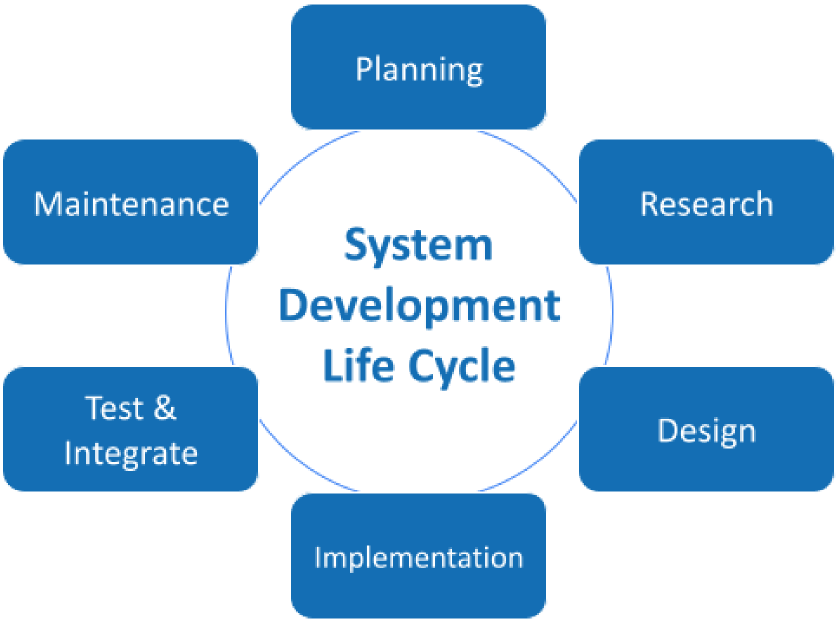


Figure 5.1: System Development Life-Cycle

5.1 | Planning I

TASI has been contracted by CNMI to create an infrastructure that will allow for data analytics on Protected Health Information (PHI). To achieve this, TASI will provide a Platform as a Service (PaaS) solution, by hosting SAS services on on-premises hardware, configured for multi-tenancy.

Tenants will provide the data, which will be submitted through an ETL pipeline for data migration, cleaning, and processing. Once the data has been processed, tenants may perform data analytics using advanced algorithms in SAS programming language.

Due to SAS being a time sensitive project, the initial deployment will have SAS suites and VMs installed on existing hardware, with plans to migrate the infrastructure to newly acquired hardware in the future.

MLA I will expect 4 tenants:

1. Commonwealth of the Northern Mariana Islands (CNMI)
2. All-Payer Claims Database (APCD)
3. Centers for Medicare & Medicaid Services (CMA)
4. Med-Quest

5.2 | Requirement of Analysis I

The Requirement Analysis phase is a crucial component in developing a robust SAS infrastructure using the SDLC framework. This phase involves gathering and analyzing the specific requirements for the project, including pre-installation checklists and EEC sizing requirements. In this phase, ongoing project management tasks will be performed, such as preparing a project plan and assigning appropriate resources.

Furthermore, as part of this phase, SAS will send a Pre-Install Requirements Document to the client for completion, and both parties will ensure environmental readiness for installation by reviewing the completed document.

Additionally, SAS will send a billable work hours log for verification based on the project plan. This subsection will provide a detailed overview of the pre-installation checklist and EEC sizing requirements necessary for a successful implementation of the SAS infrastructure.

5.2.1 | TASI's Infrastructure

The internal infrastructure of UHTASI is designed to ensure secure and efficient environment. The process begins with an internet connection, which is routed through the UH internet. To protect the network, we have implemented both a North/South (N/S) firewall and an East/West (E/W) firewall, which serve as barriers against unauthorized access and help to safeguard our data.

For enhanced reliability, redundancy is a key aspect of our infrastructure. We have two network switches in place, ensuring that if one switch fails, the other seamlessly takes over to maintain uninterrupted connectivity.

At the core of our infrastructure, we have a Dell FX2 Enclosure. The FX2 Enclosure is a 2U rack-based server located inside a server rack at the ITS data center (ITS M01). There are four blade servers that exist within the enclosure. Each blade is a self-contained server that contains one or more CPUs, memory, storage, and other components required to run applications and services..

To facilitate data storage and retrieval, we have incorporated two SAN (Storage Area Network) network switches for redundancy. These switches provide a dedicated network infrastructure for our storage, ensuring fast and reliable access. In addition, UHTASI has installed an expansion slot into the SAN to increase the overall capacity of the SAN's storage.

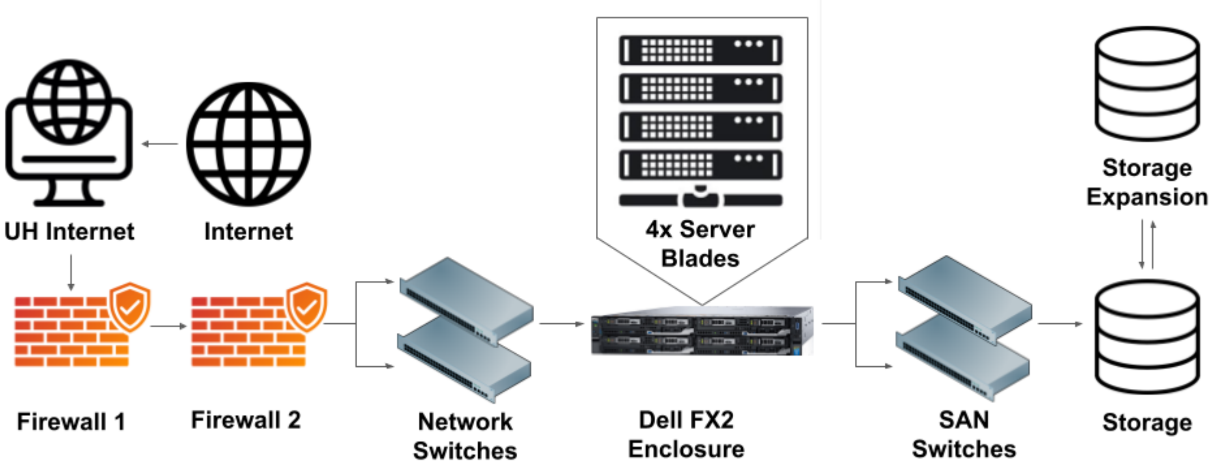


Figure 5.2: UHTASI On-Premise Infrastructure

Despite each blade in the FX2 Enclosure already being allocated to other TASI projects, the unused resources will be logically separated to establish a multi-tenant environment that can facilitate SAS technologies.

5.2.2 | Multi-Tenancy Configuration Plan: VM Location

Server Name	Function	Type	Site	Physical Server
DC1	LDAP Host1	VM	ITS M01	FX2Blade4
DC2	LDAP Host2	VM	ITS M01	FX2Blade1
SAS 9.4 Server	SAS Infrastructure Server	VM	ITS M01	FX2Blade2
SAS DMA	SAS Data Management Advanced	VM	ITS M01	NaN
SAS Ansible	Ansible	VM	ITS M01	NaN
SAS PRT	SAS Programming Run-Time	VM	ITS M01	NaN
SAS SL	SAS Service Layer	VM	ITS M01	NaN
Provider CC1	Provider Primary CAS Controller	VM	ITS M01	NaN
Provider CC1	Provider Backup CAS Controller	VM	ITS M01	NaN
R1 CC1	Research 1 CAS Controller 1	VM	ITS M01	NaN
R1 CC2	Research 1 CAS Controller 2	VM	ITS M01	NaN
R1 W1	Research 1 CAS Worker 1	VM	ITS M01	NaN
R1 W2	Research 1 CAS Worker 2	VM	ITS M01	NaN
R1 W3	Research 1 CAS Worker 3	VM	ITS M01	NaN
R2 CC1	Research 2 CAS Controller 1	VM	ITS M01	NaN
R2 CC2	Research 2 CAS Controller 2	VM	ITS M01	NaN
R3 W1	Research 3 CAS Worker 1	VM	ITS M01	NaN
R3 W2	Research 3 CAS Worker 2	VM	ITS M01	NaN
R3 W3	Research 3 CAS Worker 3	VM	ITS M01	NaN
R4 CC1	Research 4 Primary CAS Controller	VM	ITS M01	NaN
R4 CC2	Research 4 Backup CAS Controller	VM	ITS M01	NaN
R4 W1	Research 4 CAS Worker 1	VM	ITS M01	NaN
E1 CC1	Education 1 Primary CAS Controller	VM	ITS M01	NaN
E1 CC2	Education 1 Backup CAS Controller	VM	ITS M01	NaN
E1 W1	Education 1 CAS Worker 1	VM	ITS M01	NaN
E2 CC1	Education 2 Primary CAS Controller	VM	ITS M01	NaN
E2 CC2	Education 2 Backup CAS Controller	VM	ITS M01	NaN
E2 W1	Education 2 CAS Worker 1	VM	ITS M01	NaN
E3 CC1	Education 3 Primary CAS Controller	VM	ITS M01	NaN
E3 CC2	Education 3 Backup CAS Controller	VM	ITS M01	NaN
E3 W1	Education 3 CAS Worker 1	VM	ITS M01	NaN
E4 CC1	Education 4 Primary CAS Controller	VM	ITS M01	NaN
E4 CC2	Education 4 Backup CAS Controller	VM	ITS M01	NaN
E4 W1	Education 4 CAS Worker 1	VM	ITS M01	NaN

Figure 5.3: Physical and logical locations of VMs related to SAS technologies.

5.2.3 | Multi-Tenancy Configuration Plan: Resource Allocation

Server Name	Tenant	OS	Memory (GB)	vCPU	Min Sys Storage	Storage (GB)
DC1	NaN	RHEL 8	12	4	NaN	50
DC2	NaN	RHEL 8	12	4	NaN	50
SAS 9.4 Server	NaN	RHEL 8	32	8	NaN	NaN
SAS DMA	NaN	RHEL 8	32	8	NaN	NaN
SAS Ansible	NaN	RHEL 8	16	2	NaN	NaN
SAS PRT	NaN	RHEL 8	64	6	NaN	NaN
SAS SL	NaN	RHEL 8	32	2	NaN	NaN
Provider CC1	Provider	RHEL 8	8	2	NaN	NaN
Provider CC1	Provider	RHEL 8	8	2	NaN	NaN
R1 CC1	Tenant 1	RHEL 8	16	2	NaN	NaN
R1 CC2	Tenant 1	RHEL 8	16	2	NaN	NaN
R1 W1	Tenant 1	RHEL 8	16	2	NaN	NaN
R1 W2	Tenant 1	RHEL 8	16	2	NaN	NaN
R1 W3	Tenant 1	RHEL 8	16	2	NaN	NaN
R2 CC1	Tenant 2	RHEL 8	16	2	NaN	NaN
R2 CC2	Tenant 2	RHEL 8	16	2	NaN	NaN
R3 W1	Tenant 2	RHEL 8	16	2	NaN	NaN
R3 W2	Tenant 2	RHEL 8	16	2	NaN	NaN
R3 W3	Tenant 2	RHEL 8	16	2	NaN	NaN
R4 CC1	Tenant 3	RHEL 8	8	1	NaN	NaN
R4 CC2	Tenant 3	RHEL 8	8	1	NaN	NaN
R4 W1	Tenant 3	RHEL 8	8	1	NaN	NaN
E1 CC1	Tenant 4	RHEL 8	8	1	NaN	NaN
E1 CC2	Tenant 4	RHEL 8	8	1	NaN	NaN
E1 W1	Tenant 4	RHEL 8	8	1	NaN	NaN
E2 CC1	Tenant 5	RHEL 8	8	1	NaN	NaN
E2 CC2	Tenant 5	RHEL 8	8	1	NaN	NaN
E2 W1	Tenant 5	RHEL 8	8	1	NaN	NaN
E3 CC1	Tenant 6	RHEL 8	8	1	NaN	NaN
E3 CC2	Tenant 6	RHEL 8	8	1	NaN	NaN
E3 W1	Tenant 6	RHEL 8	8	1	NaN	NaN
E4 CC1	Tenant 7	RHEL 8	8	1	NaN	NaN
E4 CC2	Tenant 7	RHEL 8	8	1	NaN	NaN
E4 W1	Tenant 7	RHEL 8	8	1	NaN	NaN

Figure 5.4: Resource requirements of VMs related to SAS technologies.

5.2.4 | EEC Sizing and Pre-Installation Checklist: File Path(s)

The full EEC Sizing and Pre-Installation Checklist(s) documents can be found in:

- PATH: \ \ .. 300 SAS Installation \ 9.4 \ EEC Sizing Results
- PATH: \ \ .. 300 SAS Installation \ SAS Viya 3.5 \ EEC Sizing Results

5.2.5 | EEC Sizing: SAS 9.4 (Summarized)

This document provides sizing guidance for SAS Office Analytics/Data Management Advanced. The estimate provided assumes a typical implementation of SAS Office Analytics/Data Management Advanced and does not take into account any additional workloads or components that may be added. The estimate is based on a preferred hardware vendor with a given performance characteristic. It is recommended that the environment be closely monitored and scaled to support the required workloads to meet the business objectives.

1. Hardware and Operation System Assumptions:

Tier	Cores\RAM
SAS Metadata Server	2 cores with 16GB RAM (8 GB RAM per core minimum)
SAS Compute Server	6 to 8 cores with 48 to 64GB RAM (8 GB RAM per core minimum)
SAS Mid-Tier Server (Web-App Server)	2 cores with 24GB RAM (24 GB RAM per server minimum)

Figure 5.5: Hardware estimate for SAS 9.4: SAS Data Management Advanced.

   - This response is based on Intel Xeon E5-2600v4 or Gold 6200/6300 series processor with a clock speed of at least 3.30 GHz running Windows Server 2019, 64 bit operating system.
   - Core counts are guidelines only. These requirements may vary depending on the solutions installed or the number of users/sessions supported in accordance with Operating System Guidelines and SAS recommendations, page file space should be set to 1.5 to 2 times the amount of physical memory. The machines should be configured for maximum memory bandwidth; this will be dependent on the actual processors/machines selected.

2. SAS Environment and Configuration Assumptions:
   - SAS tends to be I/O intensive. Consider the peak I/O throughput requirements of their system and work with their storage provider to ensure that the storage environment can provide the level of I/O required. A significant percentage of “performance problems” reported to SAS Technical Support can be directly attributed to insufficient levels of I/O throughput.
   - Recommended I/O throughput rates for the SAS Data and SAS WORK file systems are as follows: for permanent SAS data files, your application throughput requirements may dictate a minimum I/O throughput rate of 100-150 MBs/sec per core, minimum, in the system. Reads and writes to the file system will occur during the ETL process. Chronic and heavy reads and writes are common for the SAS WORK file system.
   - Depending on the architecture and deployment, multiple compute tiers need access to a common data area. This may require the use of a centralized storage mechanism such as a Clustered File System (CFS).

This sizing estimate is based on a combination of guidelines provided by SAS R&D, SAS Product Management, test data, and field experience. Our best practice is to provide the topology as developed by R&D and try to provide as unified a presentation of the requirements as possible. When questions on deployment arises, the Sizing team defers to the account team.

5.2.6 | EEC Sizing: SAS Viya 3.5 (Summarized)

This document provides sizing guidance for SAS Viya 3.5. The estimate is not a performance benchmark and does not provide any performance guarantee. The University of Hawai'i at Manoa is responsible for all costs associated with procuring any hardware. This estimate assumes that appropriate data management activities will happen outside of SAS In Memory, and resources for data management activities are not included in this exercise.

1. Hardware and Resource Assumptions:

Resource Type	Resource Count
# of Servers	5 (4 CAS Worker Nodes + 1 CAS Controller Node)
CPU per server	CAS Worker Node: 2 x 8 cores Intel Xeon Gold 6234 processors (3.3 GHz) CAS Controller Node: 1 x 8 cores Intel Xeon Gold 6234 processors (3.3 GHz)
Total cores	72
Memory Clock Speed	2933 MHz
RAM per node	CAS Worker Node: 192 GB CAS Controller Node: 92 GB
Operating System	Red Hat Enterprise Linux
NIC	10 GbE
SAS Version	VIYA 3.5
Local Disk per node	2 x 480 GB SSD

Figure 5.6: Hardware estimate for SAS Viya 3.5.

- This response is based on the Dell servers with Intel Xeon processors which assumes uncompressed data.
- Additional Recommendations: Server power settings need to be set to maximum, hyper-threading should be enabled for all production CPU's, storage drives should be SSD's instead of HDD's.
- Two additional servers are configured:
  - [a] SAS Programming Runtime Environment (SPRE) is the environment where SAS programs are executed. (4 cores, 96 GB RAM, 2 x 480 GB SSD).
  - [b] Dev/Test is a sandbox server to test the development environment before production (16 cores, 192 GB RAM, 2 x 480 GB SSD).

This sizing estimate is based on a combination of guidelines provided by SAS R&D, SAS Product Management and test data. Changes to the workload (in either number of sessions or data volumes), operating system, or preferred vendor or chipset may render this sizing as void. In the event of changes, the SAS Account Team should resubmit the questionnaire with the needed updates for reprocessing.

5.2.7 | Pre-Installation Requirements Document: File Path

The full Pre-Installation Requirements document can be found in:

- PATH: \\..300 SAS Installation \ SAS Viya 3.5 \ [Fill Me]

5.2.8 | Pre-Installation Requirements Document: SAS Viya 3.5 (Summarized)

The Pre-Installation Requirements Document (PIRD) is an extensive spreadsheet that contains installation details for SAS Viya 3.5. This document encompasses the entire configuration plan, including system requirements, file systems, networking and firewall settings, authentication and encryption protocols, as well as service account requirements.

1. PIRD Form Identification

Form Identification	Metadata
Date:	4/24/2023
PIRD Template Version	Version 1.0
Last Updated	5/9/2023 3:43PM
Customer Name	UHTASI
Customer Contact	NaN
SAS Project Manager	Eric Kaiser
SAS Architect	Chauncey Cleveland
SAS Platform Type	Version
Project Phas	Planning

Figure 5.7: Form Identification for SAS Viya 3.5.

2. Installation Information
   - This section focuses on the build and server details, including server names and the resource allocation needed by UHTASI for each server. Refer to 5.3 and 5.4 for more information.
3. Instructions
   - Customer Instructions: Customers are expected to complete the column fields named Output from Client, Client Provided, and Notes.
   - Architect Instructions: Architects are expected to complete the PIRD information worksheet by completing the column fields named SAS Reviewed, Output from Client, Client Provided, and Notes.
4. General System Requirements

General System Requirements	Metadata
System Document Guide	Link
Operating System Support	· RHEL 7.1 - 7.x · RHEL 8.2 - 8.x · Oracle Linux not supported
Operating System Packages and Compliance	· lilbXp & libXmu · numactl package & X11/Xmotif (GUI) · glibc-2.17-107.el7
Key Deployment Information	· Deployed using RPM packages via Ansible config tool · Needs to connect to an LDAP server for authentication
Server(s)	· All servers should be dedicated to SAS Viya · All servers should have identical CPU and Memory · Needs to connect to an LDAP server for authentication
CPU Guidelines	· Intel Xeon Chipset @ 2.6GHZ · Minimum: 4 cores · Recommended: 16 cores
Memory Guidelines	· 16GB of RAM per core @ 1600MHz · 64-96GB of RAM per machine

Figure 5.8: General System Requirements I

General System Requirements	Metadata
Special Considerations for CSP's	· To consult with a SAS sizing expert: send an email to contactcenter@sas.com
I/O Configuration	· Usage Note 51660 to test SAS throughput for file systems
Visual Interfaces	· VI's require a connection to identity provider (LDAP) · Bind: (1) anonymously, (2) with a specific binding account · Recommended to exclude account from password change · Recommended to exclude account from account locking policies
Network Considerations	· Recommended 10GB ethernet connection · Servers should have static hostnames and static IP addresses
DNS and DNS Alias	· Names need to be resolvable by all hosts in SAS · All hosts must reside in same DNS domain and sub domain
Inbound Access	· Servers hosting Viya should be accessed through internal network
Outbound Access	· Servers hosting Viya should have internet access, directly or proxy · Must configure YUM and CURL if proxy is in front of Viya servers
Firewalls	· Configure firewall to allow internal SAS Viya traffic flow
Ports	· All TCP ports (in & out) should be open between servers · Ports 80, 443, and possibly 5570, 17551, should be open · Other ports need to be opened for data sources · Refer to Deployment Guide for the complete list

Figure 5.9: General System Requirements II

5. Viya File System
6. Networking and Firewall
   - Servers should have **static hostnames** and **static IP addresses** as any future changes in hostname or IP will break the environment. Refer to 5.9 for more information on network.
   - RPM Package Downloads
     - https://ses.sas-download/
     - https://bwp1.ses.sas.download/
     - https://bwp2.ses.sas.download/
     - https://sesbw.sas.download/
   - All TCP ports should be open both ways (inbound and outbound), including single-machine environments such that the machine can connect to itself.
     - SAS Web Server (Apache HTTPD) - 443
     - CAS Client Connections (Python/R Clients to CAS Controller) - 5570
     - SAS/Connect (SAS Programming Runtime Environment) - 17551/17541
     - SAS Workspace Server - 8591 (Enterprise Guide 8.2+)
7. Multi-Tenancy
   - UHTASI has opted in for a multi-tenant environment in the initial deployment of SAS Viya.
8. Authentication (LDAP, AD, KRB5)
   - In SAS Viya 3.5, the visual interfaces require a connection to an identity provider. Binding to the LDAP identity provider can be done in two ways: anonymously or with a specific ”binding account.”

- When using a binding account, there are some important considerations regarding UserDN and Password:
    - □ The UserDN and Password will be stored in the Viya environment and used for authorization and identity mechanisms.
    - □ Any regular LDAP account can be used for this purpose.
    - □ If there are any changes to the UserDN or Password, the stored credentials must be updated in the environment to minimize downtime.
    - □ If the binding account gets locked or if the password expires without being changed, it will prevent all users from logging into the environment.
    - □ To avoid this, it is recommended to exclude the binding account from any password change or account locking policy.
- The required information for the binding account needs to be provided. If an anonymous bind is desired, ”none” can be entered in the UserDN and Password fields.
- Note that when using LDAPS (LDAP over SSL), the microservices in SAS Viya must trust the signer of the certificate to establish a secure connection.

9. Encryption SSL
    - SAS Viya 3.5 Requires the following for TLS Certificates.
        - □ BASE64 - X509 Server Identities Certificate
            - ○ The SAN of this cert should contain the FQDN of the Web Server Host(s) and all aliases.
        - □ BASE64 - RSA Private Key
        - □ BASE64 - X509 Certificate Authority Chain (root and any intermediate signers)
        - □ HA of CAS / Web servers requires a certificate to be used for each
            - ○ For example - if 2 CAS Controllers are deployed they should each leverage the same certificate which should contain both FQDNs as SANs.
10. Service Account Requirements

Service Account	Required Name	Required Characteristics
SAS Viya Deployment Account	No (e.g., viyadep)	· SUDO rights to sas, cas, and root · SSH to all Viya hosts from Ansible CTRLR · UID and GID on all SAS Viya Hosts · Either local or domain account
SAS Viya Installation User	Yes: sas	· Primary group sas · SSH to all Viya hosts from Ansible CTRLR · Non-expiring password policy · Recommended local user
SAS Viya CAS Owner	Yes: cas	· Primary group sas · SSH to all Viya hosts from Ansible CTRLR · Non-expiring password policy · Recommended local user
SAS Viya LDAP Bind Account	No (e.g., viya_bind_ldap)	· MUST be an LDAP or Domain Account · Non-expiring password policy · Normal read rights to LDAP
SAS Viya RabbitMQ Owner	Yes: sasrabbitmq	· Local Account. Created by installation
Postgres Owner	Yes: postgres	· Local Account. Created by installation

Figure 5.10: SAS Viya 3.5 Service Account Requirements

Account/Group	Required Name	Required Characteristics
Tenant Admin Account	No (e.g., acmeadmin, etc)	· Primary group sas · UID and GID on all SAS Viya Hosts · No password assigned · Non-expiring password policy · Recommended domain group
Tenant Admin Group	No (e.g., acmeadmgroup)	· Can either be local or domain group · Recommended domain group
SAS Provider End Users	N A (end users)	NaN
Tenant User Group	No (e.g., acmeusergroup)	· Can either be local or domain group · Recommended domain group

Figure 5.11: SAS Viya 3.5 Multitenant User and Group Requirements

11. Deployment Tools (Ansible)
    - Before installing SAS Viya, UHTASI is required to install Ansible, a software tool that facilitates infrastructure as code for managing IT environments. The following section offers step-by-step instructions on how to install Ansible on your Linux machine.
    - e.g.: *$sudo yum install -y ansible*.
12. Pre-Installation Playbook (Ansible)
    - To streamline and automate the setup process for Viva deployment, SAS has developed an Ansible playbook as part of the Viya Administration Resources Kit (Viya-ARK). By leveraging this playbook, customers can save time and effort when performing and verifying the necessary pre-requisites outlined in the Deployment Guide.
    - The playbook is readily accessible on Github.
13. Server Requirements Checklist
    - This section is a comprehensive log report that serves as a record of each installation requirement check and step. This systematic approach allows us to track the progress of the installation accurately.
    - Whilst the previous sections (3-12) provide general information, this section offers more detail on the context of each step, suggested validation commands to run, and the expected result after completing the step.

Item	Reviewed	Validation Command	Expected Result
libpng12 Package	Yes	*$ rpm -q libpng12*	libpng12-1.2.50-7.el7_2.x86_64

Figure 5.12: Server Requirements Checklist (OS System Package Example)

14. Datasources
    - To ensure compatibility, UHTASI is required to specify the type of data source that SAS Viya will be accessing. By declaring the data source, the necessary checks can be performed to ensure that the versions of the data sources align with the supported configurations for SAS Viya.
    - ***May 15, 2023:*** *UHTASI's data source is fully compatible with the latest versions of SAS Viya*.
15. Create Mirror Repository
    - This ”mirror” term refers to the fact that we are building a copy of the original SAS Packages Repository. Then, the deployment can download the SAS packages from this copied repository instead of using the SAS Hosted one. This section includes instructions on how to configure a mirror repository.

- Reasons for using a mirror repository include:
    - □ SAS Viya servers do not have internet access
    - □ SAS Viya deployment in SUSE Linux environments
    - □ Multi-tenancy or multiple CAS Servers
    - □ Mirror's are static that do not dynamically update to the latest versions of packages
    - □ Reduce the security risk exposure of internet access

5.3 | Design I

The Design phase is a critical step in implementing a successful SAS infrastructure. During this phase, the technical specifications and architecture of the system are defined, and the appropriate hardware and software components are selected. This phase also includes creating a deployment plan, which outlines the steps for installing and configuring the system.

5.3.1 | Design Principles

TASI will design a multi-tenant infrastructure that will accommodate all the VMs specified in the multi-tenancy configuration plan in Section 5.2.2. The design should incorporate the architectural best practices of a Well-Architected Framework, which is a set design principles for running and designing workloads.

1. **Operational Excellence**:
    - TASI must make frequent, small, and reversible changes to hardware and software to minimize disruption to the production stage and allow for quick reversibility in case of issues.
    - TASI must constantly refine operation procedures, setting up regular game days to validate that all procedures are effective and efficient.
    - TASI must anticipate for failure by testing failure scenarios and response procedures of servers, VMs, and SAS components.
    - TASI must learn from all operational failures and share what is learned across the organization.
2. **Reliability**: A reliable application must be designed for failure by supporting high availability and disaster recovery principles.
    - CAS controller and CAS backup controller nodes must exist on separate hardware to support automated failover in the case of unexpected downtime.
    - TASI's on-premises infrastructure must have sufficient resources (e.g., compute, memory, storage) beyond the minimum requirement for supporting SAS technologies.
    - All data that exists in volatile memory or storage should have regular backups. SAS loads data to be analyzed into non-volatile memory.
    - All VMs should be scheduled for incremental backups and retention policies.
3. **Security**: A secure application must be designed for confidentiality, integrity, and availability, whilst also adhering to new security standards such as authorization, encryption, monitoring, and auditing.
    - Any data that TASI intends to store or process must comply with regulations and industry standards from HIPAA, UHM, RCUH, and other relevant compliance standards for protected health information.
    - TASI must consider the principle of least privilege when designing an LDAP directory to support multi-tenancy.
    - Data Governance: **Fill me on May 24**
    - HIPPA compliance standards require data to be encrypted at-rest and in-transit. Data that is loaded in non-volatile memory for SAS, does not have to be encrypted.

4. **Performance Efficiency**: An efficient application has the ability to use computing resources efficiently to meet system requirements, and maintains that efficiency as demand and technology changes.
   - CAS worker nodes must be configured for MPP mode, where possible.
   - The SAS environment must be designed on infrastructure with ample resources beyond the minimum requirements to prevent potential bottlenecks as demand scales up.
   - TASI must ensure that the servers have sufficient resources beyond the minimum requirements for VMs, to prevent potential bottlenecks when demand increases.
   - As the infrastructure grows, TASI must consider load balancing applications for user traffic across multiple servers.
   - To prepare for MLA II (VM Migration), TASI must ensure that the initial deployment is loosley coupled for scalability. This involves designing the infrastructure to be elastic, so it can handle sudden spikes in demand without compromising performance or availability.
5. **Cost Optimization**:
   - TASI will not have true cost optimization as SAS services are purchased through CAPEX.

5.3.2 | IAM Design

RCUH and UHTASI.

if uh then uh makes their own active directory. (check asana cause athena is working on that information with its

if tasi, we have to design our own LDAP system which is goin gto suck

[user] -¿ [vpn] -¿ [connection is approved by LDAP]

5.3.3 | Multi-Tenancy

The initial deployment of MLA will involve the installation of SAS 9.4 and SAS Viya 3.5 on existing infrastructure. The deployment configuration for each tenant will be tailored to meet their individual requirements.

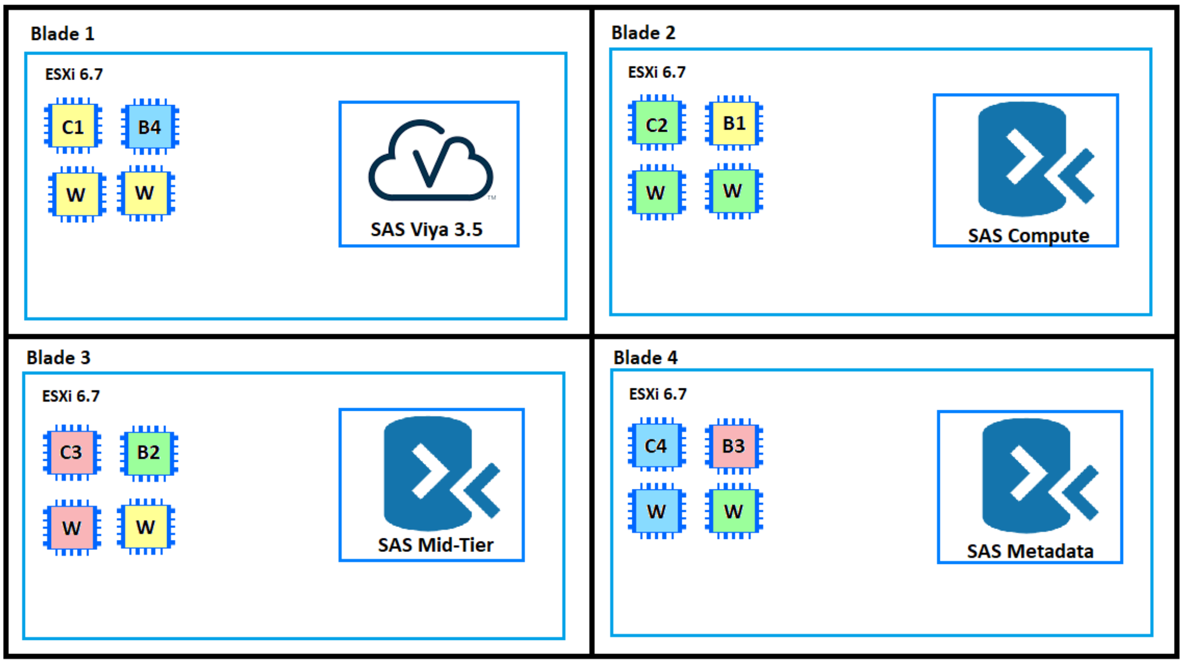


Figure 5.13: Multi-Tenant Deployment (needs VISIO)

To maximize resource efficiency, CAS nodes will be evenly distributed across each blade, where a blade will consist of one controller, one backup controller, and two workers. The controller and backup controller, configured on the same system, will belong to separate tenants. The workers will also belong to separate tenants but each blade will have at least one related controller and worker per system.

Subsequently, four additional VMs will be created to support the installation of SAS Viya 3.5 and SAS DMA. SAS Viya 3.5 will be installed as software on top of a RHEL 3.7X VM instance, in Blade 1. SAS DMA consists of three software components that will installed as software on top of Windows Server 2019 VM instances, in Blades' 2, 3, and 4.

5.4 | Implementation I

April 18, 2023: SAS 9.4 is installed first on TASI system. It is configured this way on blade 1.

5.5 | Testing & Integration I

5.6 | Operations & Maintenance I

6 | Hyper-Converged Infrastructure (HCI)

HCI, or Hyper-Converged Infrastructure, is a software-defined, unified system that combines the traditional elements of IT infrastructure (e.g., compute, networking, management, storage) with virtualization, simplifying infrastructure, reducing costs, and increasing scalability and flexibility. In a traditional IT Infrastructure, servers, storage networks, and storage systems are physically separated as stand alone hardware devices (e.g., servers, network switches, disk arrays). Consolidating these components into a single, integrated system simplifies the management, deployment, configuration, and maintenance of your IT Infrastructure.

The benefits of an HCI environment include:

- Scalability: Designed to scale out by adding additional nodes on-demand to your system.
- Efficiency: Improve resource utilization by using or eliminating idle storage capacity.
- Agility: Quickly deploy new applications and workloads without extensive planning across systems.
- Data Protection: Integrated backup and disaster recovery.
- Reduced Hardware Costs: Reduce the amount of hardware required reducing CAPEX[14]/OPEX[15] costs.

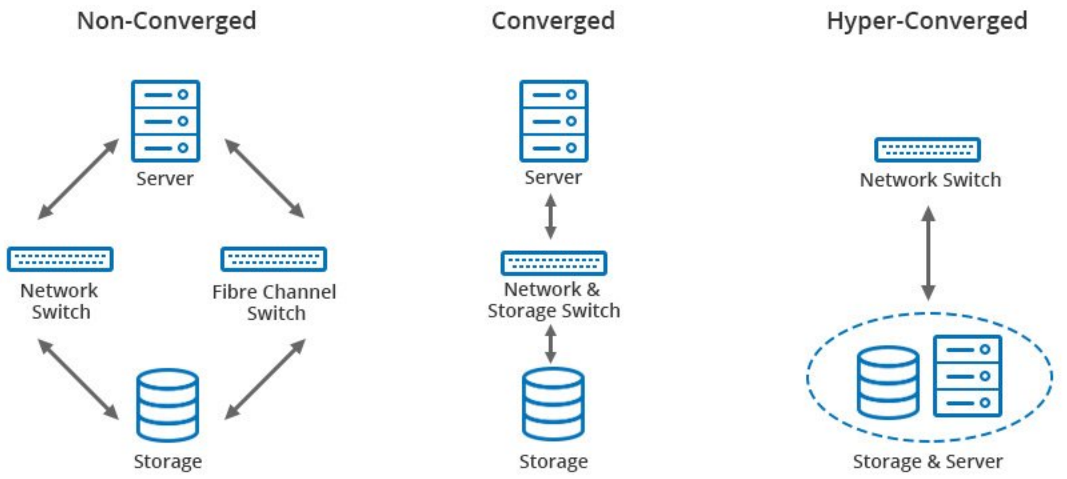


Figure 6.1: Types of IT Infrastructures (STOLEN EXAMPLE)

In HCI, multiple servers or nodes are combined to create a cluster. These nodes share their computing and storage resources with each other to create a multi-purpose integrated system. The design of your HCI cluster will depend on your specific needs and requirements.

The software that powers HCI also includes a management layer, which automates tasks like resource provisioning, data migration, and load balancing. This layer abstracts the hardware, making it easier to manage and deploy your IT infrastructure. Overall, HCI is a powerful and flexible solution that can help organizations streamline their IT operations, reduce costs, and improve efficiency.

[14] Capital expenditure is the cost a business incurs to acquire assets that will provide benefits beyond the current year.
[15] Operating expenses refer to the money a company spends to run day-to-day operations.

7 | Massively Learning Activities II - Migration Deployment

TASI has been contracted by CNMI to create an infrastructure that allows for data analytics on Protected Health Information (PHI). This infrastructure will initially be hosted on-premises, with plans to move towards a hybrid solution in the future. To achieve this, we will be providing a Platform as a Service (PaaS) solution, by hosting SAS Viya services on our own hardware and allowing tenants to access and utilize the platform for their own analytics applications.

The tenants, including APCD, CMNI, CMA, Criminal Justice, and several Education environments, will provide the necessary data, which will be submitted to an ETL data pipeline for processing before being sent to SAS on-prem servers. Once the data has been processed, tenants may perform data analytics using advanced algorithms in SAS programming language.

To ensure secure operations, we will configure the security relationships between the software, hardware, and tenants using LDAP, security groups, encryption and other related tools. Our goal is to architect a high-performance infrastructure that allows for advanced data analytics while maintaining the confidentiality and security of PHI.

Due to SAS being a time sensitive project, the initial deployment will have SAS suites and VMs installed on existing hardware, with plans to migrate the infrastructure to newly acquired hardware in the future.

The final and completed deployment of SAS Viya 3.5 will expect a total of 8 tenants:

1. Commonwealth of the Northern Mariana Islands (CNMI)
2. All-Payer Claims Database (APCD)
3. Centers for Medicare & Medicaid Services (CMA)
4. Med-Quest
5. University Education 1
6. University Education 2
7. University Education 3
8. University Education 4

7.1 | Planning II

The System Development Lifecycle (SDLC) is a project management model that defines different stages that are necessary to bring a project from conception to deployment and later maintenance. The SDLC model consists of several phases, which typically include requirements gathering, design, development, testing, deployment, and maintenance. The specific activities within each phase may vary depending on the project and the organization, but the basic principles are the same. The SDLC model is a flexible framework that can be adapted to suit the needs of different projects and organizations. It provides a systematic approach to software development that helps ensure that software is built efficiently, effectively, and with minimal risk.

Massively Learning Activities will follow a similar variation to the SDLC project management model where each SDLC stage will correspond to a subsection in this chapter.

7.2 | Required of Analysis II

- Requirement of Analysis
- Deployment Design

7.3 | Design II

We will be using VMotion to migrate virtual machines.

7.4 | Implementation II

7.5 | Testing & Integration II

7.6 | Operations & Maintenance II

A | Appendix A title

Acronym	Meaning
ABAC	Attribute-Based Access Control
APCD	All-Payer Claims Database
CAS	Cloud Analytic Services
CMA	Centers for Medicare & Medicaid Services
CNMI	Commonwealth of the Northern Mariana Islands
CPU	Central Processing Unit
CUR	Central User Repository
DMA	Data Management Advanced
EHR	Electronic Health Record
ETL	Extract, Transform, Load
HIPAA	Health Insurance Portability and Accountability Act
HIT	Health Information Technology
IAM	Identity and Access Management
ICA	Intergovernmental Cooperative Agreement
ICT	Information and Communication Technology
LDAP	Lightweight Directory Access Protocol
MPP	Massively Parallel Processing
PHIDC	Pacific Health Informatics Data Center
RAM	Random Access Memory
RBAC	Role-Based Access Control
RCUH	The Research Corporation of the University of Hawaii
RPO	Recovery Point Objective
RTO	Recovery Time Objective
SAS	Statistical Analytic Services
SDLC	System Development Life-cycle
SMA	State Medicaid Agency
SMP	Symmetric Multi-Processing
SSL	Secure Sockets Layer
SSO	Single Sign-On
SSRI	Social Science Research Institute
TASI	Telecommunications and Social Informatics Program
TLS	Transport Layer Security
VM	Virtual Machine

Figure A.1: A list of all relevant acronyms in this document.