TASI
TELECOMMUNICATIONS AND SOCIAL INFORMATICS RESEARCH PROGRAM

# Massively Learning Activities

**In Woo Park**

---

Master's Capstone Project

Master's of Science in Computer Science

University of Hawaii at Manoa

Under the Supervision of:

Professor Edoardo Biagioni

July 1, 2023

## Abstract

This paper provides an overview of UHTASI's comprehensive project with SAS technologies, which encompasses various aspects and involves partnerships with multiple agencies in the Pacific region. The project's primary objective is to establish an infrastructure for deploying SAS technologies to perform data analytics on sensitive PHI data, including healthcare claims and criminal justice records. Recognizing the urgency of the project, UHTASI has made concerted efforts to expedite the deployment of SAS technologies, focusing on their multi-tenancy capabilities and utilizing existing on-premises hardware.

Following a successful initial deployment, UHTASI plans to acquire additional hardware and migrate the existing SAS infrastructure to a new hyper-converged infrastructure (HCI) developed by UHTASI. Once the migration to HCI is complete, UHTASI aims to replicate the HCI environment and incorporate four additional tenants, enabling faculty and students at the University to access SAS technologies for research and educational purposes. Notably, the HCI instance for the University will be segregated from the original HCI, ensuring separate and secure environments.

This paper highlights the progressive steps taken by UHTASI in deploying SAS technologies, emphasizing the efficient utilization of resources and collaborative partnerships. The successful implementation of this project will contribute to improved data analytics capabilities and facilitate academic engagement with SAS technologies in the University setting.

# Table of Contents

# 1 | Introduction

In today's data-driven world, the effective utilization of technology and information systems is paramount to gain valuable insights and make informed decisions. Within the realm of healthcare, the ability to analyze vast amounts of data, such as Protected Health Information (PHI), can lead to improved patient care, enhanced policy planning, and more efficient healthcare systems. By harnessing the power of SAS analytics, this project aims to revolutionize data management and analytics capabilities, enabling researchers, practitioners, and policymakers to gain valuable insights and drive innovation in the healthcare domain.

## 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 [20].

**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 40 Hours - Subject To Change)
   - 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 140* Hours - Subject To Change)
   - During the deployment phase, TASI will receive the installation of several SAS products:
     - SAS Visual Analytics
     - SAS Visual Statistics
     - SAS Visual Data Mining and Machine Learning
     - SAS Visual Forecasting
     - SAS In-Memory Statistics
     - Viya Platform
     - Model Manager
     - SAS IML (Interactive Matrix Language)
     - SAS QC (Quality Control)
     - SAS Econometrics
     - Data Preparation
     - SAS/ACCESS Engines
     - Visual Analytics Add-In For Office
     - Base SAS
     - Data Management Advanced Server
     - Enterprise Guide
     - SAS/ACCESS to MS

# 2 | 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 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.

## 2.1 | Data Compliance

- HIPAA Compliance [22]
- UHM Compliance
- RCUH Compliance
- TASI Compliance
- State of Hawaii Compliance

## 2.2 | 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 [2].

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.

### 2.2.1 | 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).

### 2.2.2 | 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).

### 2.2.3 | 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.

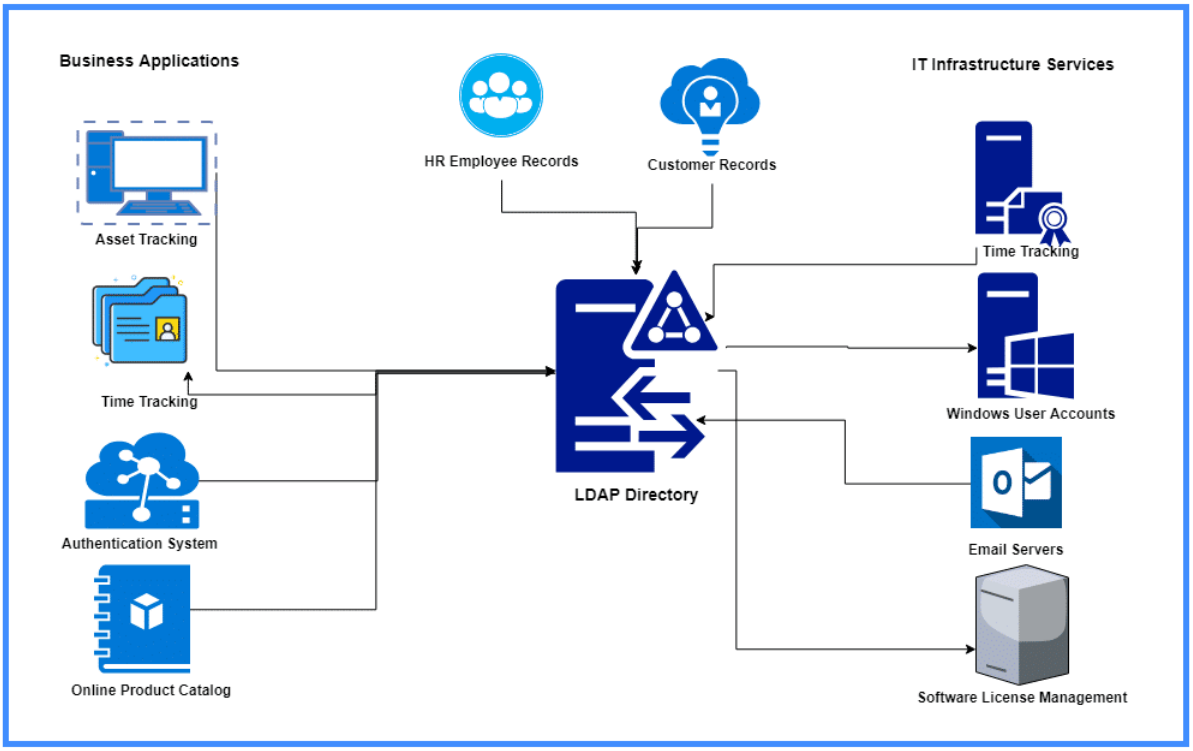


**Figure 2.1:** Lightweight Directory Access Protocol [19]

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.

### 2.2.4 | User Management

User Management is the component of IAM that covers the creation and maintenance of user accounts, account identity, and account privileges [4].

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 2.2.1).

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.

## 2.3 | 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 [11].

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, tape, 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.

## 2.4 | 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.

## 2.5 | Vulnerability Assessment

A Vulnerability Assessment is the process of identifying vulnerabilities in a system to assess potential security risks [8].

It involves scanning the system for known weaknesses, misconfigurations, or software vulnerabilities that could be exploited by bad actors. By conducting regular vulnerability assessments, organizations can proactively identify areas of weakness and take appropriate measures to mitigate potential threats.

The assessment typically follows these steps:

1. Vulnerability Scan:
   - Utilizing automated tools to scan an organization's IT environment to identify vulnerabilities and weaknesses.
2. Vulnerability Analysis:
   - Determine the severity and impact on the system's security. Prioritize the vulnerabilities based on the level of risk.
3. Risk Assessment:
   - Prioritize the remediation efforts based on the level of risk associated with each vulnerability.
4. Remediation Planning:
   - Develop a plan to address the identified vulnerabilities.
5. Ongoing Monitoring:
   - Vulnerability assessments should be conducted on a regular basis to ensure that the system remains secure over time.

## 2.6 | Penetration Testing

Penetration Testing, also known as ethical hacking, is a simulation of real-world attacks on a system to evaluate the security of a system [8].

1. Planning:
   - Defining the penetration test objectives and identifying the target system.
2. Reconnaissance:
   - Gathering available information about a system through passive methods.
3. Vulnerability Assessment:
   - Identify and assess the vulnerabilities in a target system using automated tools.
4. Exploitation:
   - Exploit the identified vulnerabilities to gain unauthorized access, escalate privileges, or compromise sensitive data.
5. Post-Exploitation and Reporting:
   - A detailed report is generated to show the vulnerabilities discovered, how it was exploited, and how to remediate them.

# 3 | Data Governance

Data Governance, is the organizing framework that establishes internal data policies that apply to how data is gathered, stored, processed [9].

In the context of designing an infrastructure with SAS technologies, UHTASI assumes the responsibility of establishing data governance policies that adhere to HIPAA compliance standards, standardized security practices, and industry best-practices of sharing data [12]. This obligation arises from the overarching goal of MLA, which is to provide data analytic services for PHI data.

Note that this chapter does not aim to be a comprehensive handbook for the implementation of proper data governance but offers an overview of data governance methodologies. Information was gathered from the SAS conference on data governance (May, 2023).

## 3.1 | SAS Data Management Framework

The SAS Data Management Framework is a collection of groups and methodologies that offer solutions for data integration, data quality, data governance, and metadata management [21].

| Principle | Description |
|---|---|
| Program Objectives | The, ”why?”, establishes the goals and purpose of the data governance program. |
| Guiding Principles | The, ”how?, outlines the framework for making data-related decisions and ensuring consistency and alignment with organizational goals and values. |
| Decision-making Bodies | The, ”who?”, identifies the decision-making bodies or committees responsible for overseeing and making key data governance decisions, ensuring representation from relevant experts. |
| Decision Rights | The, ”what?”, allocates decision-making authority and responsibilities to clarify accountability and effective data management and control. |

**Figure 3.1:** 4 Principles of Data Governance

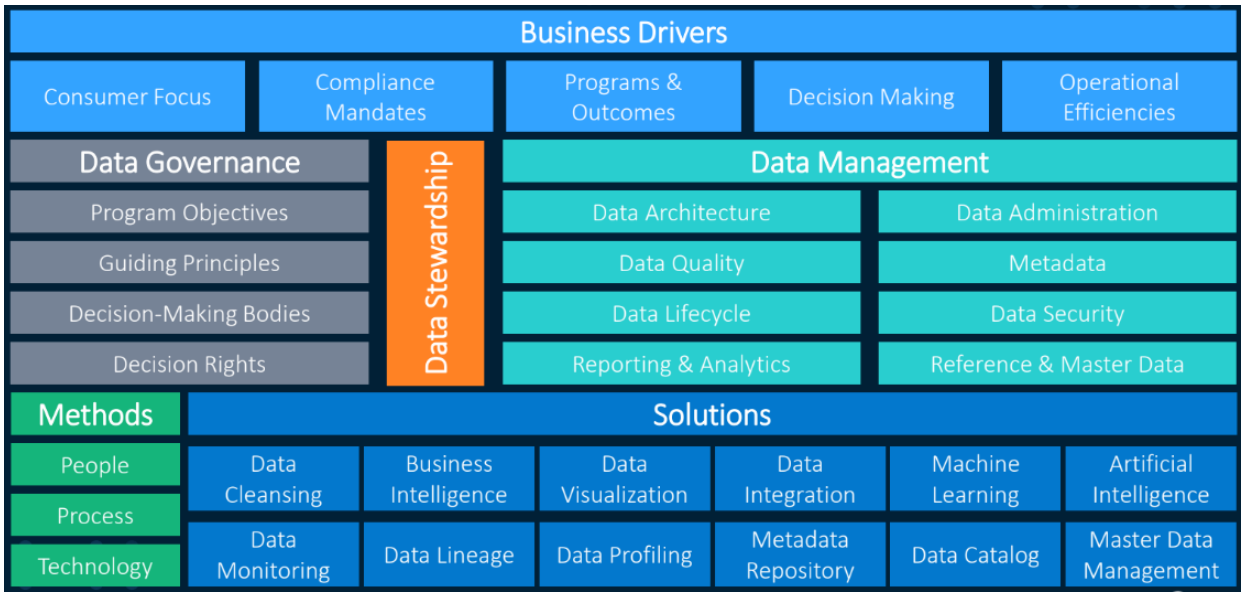


**Figure 3.2:** SAS Data Management Framework

SAS Data Management, as part of the SAS Data Management Framework, is a set of principles that cover policies and standards about data collection, storage, and processes.

| Principle | Description |
|---|---|
| Data Architecture | Models, policies, rules, standards about what data is captured, how it is stored, arranged and integrated (data analysis, data modeling, data design). |
| Data Quality | Quality of data, its integrity and how its is cleansed and/or enriched. |
| Data Lifecycle | How data is created, stored, distributed, used, maintained, archived, and disposed. |
| Reporting | Analytics & Systems that support the creation of value from data. |
| Data Administration | Day-to-day management and control of data and databases. |
| Metadata | Capture, storage, documentation, and publishing of information about enterprise data such as its description, lineage, usage, ownership, etc. |
| Data Security | How data is accessed and secured and how privacy is handled. |
| Reference & Master Data | Correct and consistent management of master and reference data. |

**Figure 3.3:** 8 Principles of Data Management

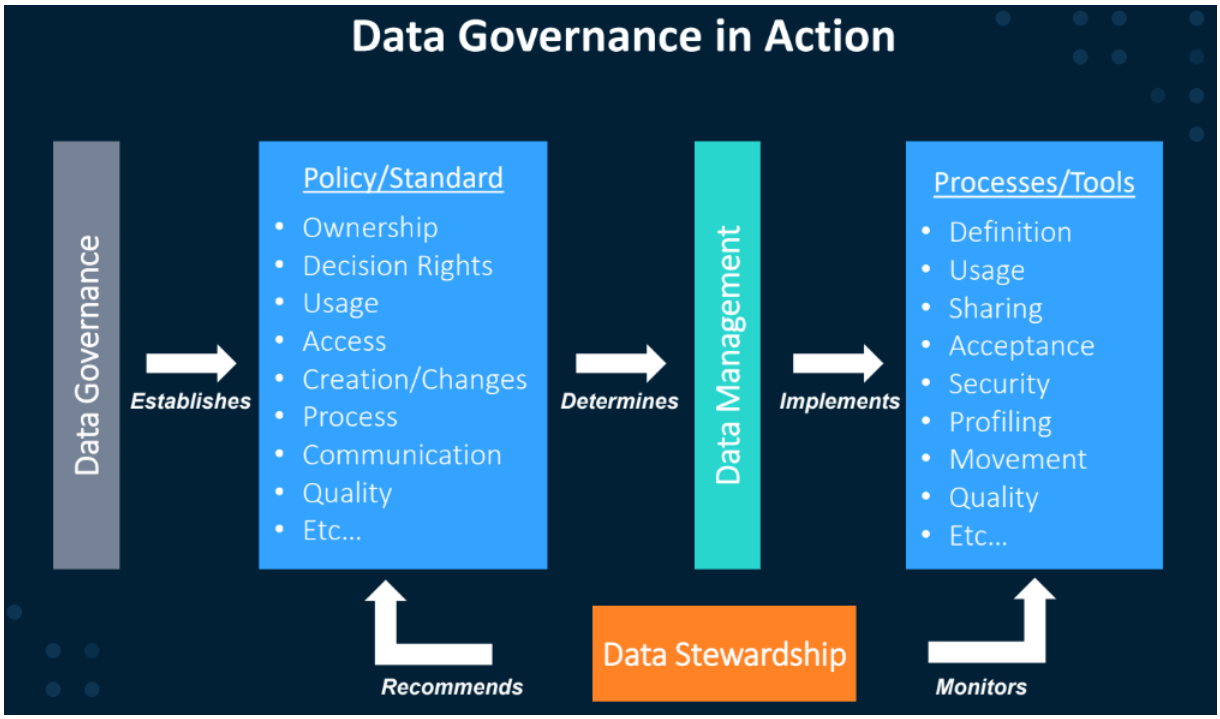


**Figure 3.4:** SAS Data Governance in Action

Engaging in data governance without achieving data management outcomes is merely an academic exercise and is bound to fail. Similarly, conducting data management without proper data governance perpetuates a culture of reliance on unreliable sources and informal knowledge sharing.

Data stewards play a vital role in actively monitoring data and advocating for organizational policies, serving as the vital link between data governance and data management activities.

## 3.2 | Data Stewardship

A data steward is someone whose role is to develop and protect the information resources for the organization and ensures the integrity of the data. However, data stewardship is not data governance, it is a critical part of data governance.

| Principle | Description |
|---|---|
| Metadata | Common definitions, Data dictionary, Data mining, Data lineage. |
| Data Quality | Continuous improvement, Root cause analysis, Measurement. |
| Architecture | Standards, Scalable solution(s). |
| Data Trust | Standards, Policies. |
| Data Integration | Transparent processes, Data ready for use. |
| People and Process | Teamwork, Facilitation, Consensus building. |
| Business Needs | Identify, Set priorities, Alignment of data with business needs. |
| Communication | Catalyst for change, Data evangelist. |

**Figure 3.5:** 8 Principles of Data Stewardship

## 3.3 | Data Lineage

Data lineage refers to the ability to trace the origin, movement, and transformation of data throughout its lifecycle.

Data lineage provides an understanding of how data is created, modified, and used within an organization. By documenting data lineage, organizations can establish a record of data flow to enable better data governance.

The specific implementation of data lineage within an application can vary depending on the tool or platform used. However, most applications will follow the general methodologies, such as metadata collection, metadata association, data flow mapping, audits, and reports.

An important aspect of data lineage is not only the ability to view the transformation of data over time, but also the ability to audit. With the audit functionality in data lineage, organizations can effectively trace and document user actions (i.e., data: view, add, delete, change). This audit trail provides a record of who has interacted with the data, when these interactions occurred, and what changes were made. By having this level of visibility, organizations can strengthen their data governance practices and gain confidence that only authorized users are accessing the data.

Reports can be generated from these audits to be used for compliance reviews.

## 3.4 | Metadata

Metadata can be defined as data about data, serving as valuable information that assists in navigating the intricate network of data within an organization and facilitates its effective utilization.

### 3.4.1 | Metadata Management

The practice of gathering, storing, and provisioning information about data assets.

Metadata plays a crucial role in various aspects of data management and analytics within an organization. It serves to increase confidence in the data by providing valuable context and understanding.

Operational efficiencies are harmonized as redundant data and processes are identified, streamlining operations. Effective communication between data creators, consumers, and IT is facilitated, fostering collaboration and alignment. This reduces the time to market by decreasing the development life cycle and enabling faster data research. Change management and impact analysis for IT become simpler, ensuring smooth transitions and minimizing disruptions.

### 3.4.2 | Types of Metadata

There are 3 types of metadata used by organizations.

| Metadata Type | Data Owner | Metadata Objective |
|---|---|---|
| Business Metadata | Business | Provide a road map to navigate in business context the complex network of data an organization has. |
| Technical Metadata | Technical | Technical characteristics of data used by IT staff to design efficient databases, queries, and applications and to reduce data duplication. |
| Operational Metadata | Technical | Describes characteristics of routine operations. |

**Figure 3.6:** Types of Metadata

1. Business Metadata Examples:
   - Business name and description, data location, data usage, business rules, access instructions, valid range of values, relationship to other data elements, data owner and steward, and data security classification.
2. Technical Metadata Examples:
   - Table name, column name, data type, data size, primary and foreign key attributes, optionality, nullability, database server identifier, and data lineage.
3. Operational Metadata Examples:
   - Schedules and batch jobs, logs of job execution, error logs, audit, balance, and control measures, SLA requirements, volume and usage statistics, backup and archival information, and daily ETL (3.10) processes.

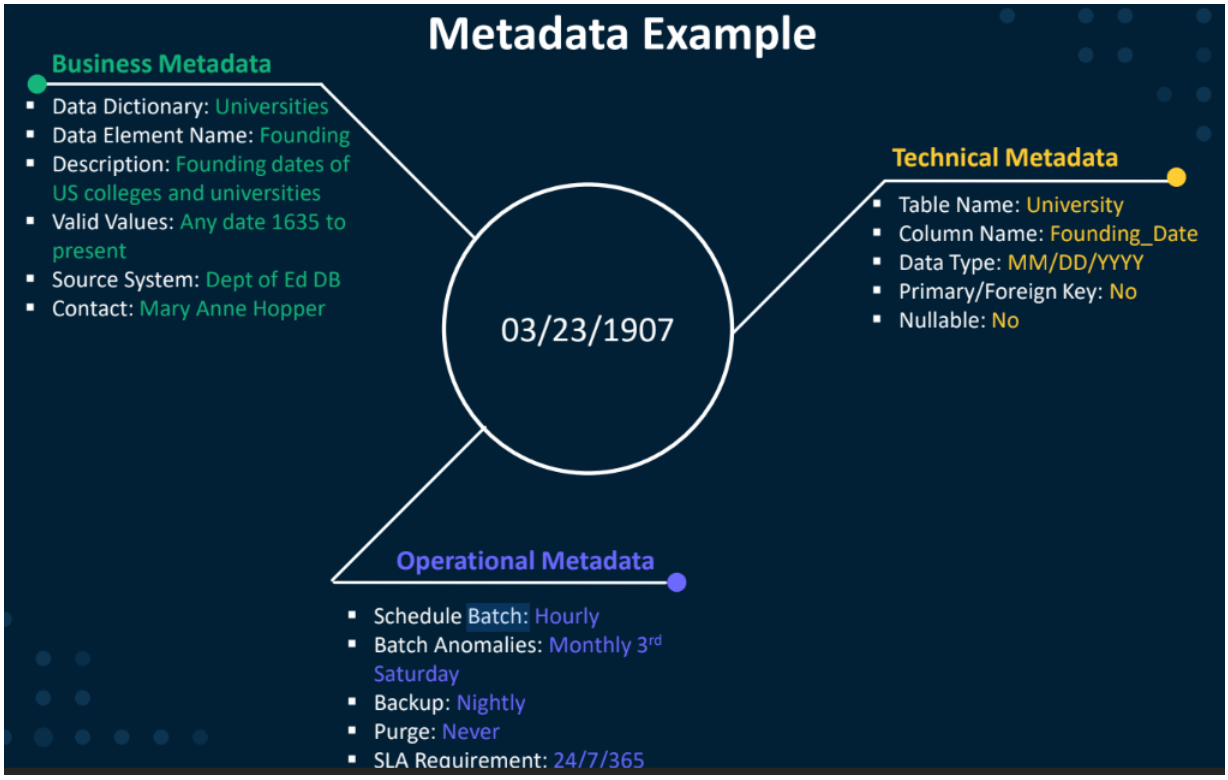


**Figure 3.7:** Example of Metadata

### 3.4.3 | Metadata Management Cycle

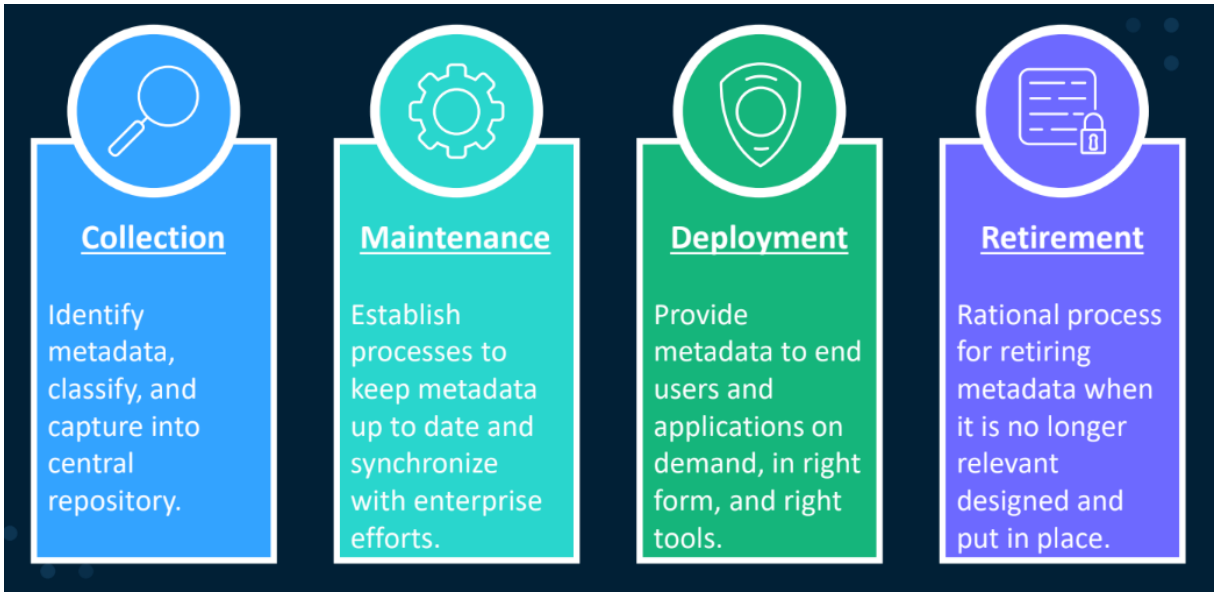


**Figure 3.8:** Metadata Management Lifecycle

## 3.5 | Data Dictionaries

A data dictionary defines the structure and contents of data sets, databases, applications, or warehouses.

A recommended approach involves using a standardized document to create a proper dictionary that provides insights into the structure and contents of data sets, databases, applications, or warehouses. Developing a data dictionary requires access to the data, which may necessitate data sharing agreements such as business association agreements for HIPAA compliance. The data dictionary acts as a reference guide that defines the various aspects of data, aiding in its management and utilization.

To begin building a data dictionary, it is recommended to start with organization critical metadata elements.

| Data Dictionary | Description |
|---|---|
| Data Element Name | Business name of element |
| Description | Business Description |
| Data Type | Physical data type (varies depending on database platform) |
| Valid Values | Valid data ranges; 1,0 for measures; valid date range |
| Source System | Source system database - may need to insert additional columns for multiple load stops (staging, DW → DM) |
| Contact | Data Owner or Data Steward contact person |

**Figure 3.9:** Key Organization Critical Metadata Elements

### 3.5.1 | Advantages of Data Dictionaries for Code Sharing

While HIPAA does not have specific provisions for code sharing, its primary focus is on safeguarding the privacy, security, and confidentiality of protected health information (PHI). To address the issue of code sharing within the context of HIPAA compliance, organizations can adopt standardization practices.

With a clear definition of data elements, their meanings, and relationships, data dictionaries promote consistency and uniformity in data usage across various systems and applications. This standardized approach enables a positive environment for the facilitation of standardized code sharing methods.

Code sharing refers to the practice of sharing snippets, modules, or programs, written in a programming language, to facilitate collaboration or knowledge exchange among developers. Therefore, when it comes to code sharing, having consistent and well-defined data elements allows developers to easily understand and use the shared code.

In addition, the code can be documented alongside data definitions to provide insights on how to use the code and the context in which the code should be applied. By adhering to defined data elements and coding conventions documented in the data dictionary, developers can produce code that follow established guidelines. This ensures that shared code integrates well across data sources and maintains a high level of quality and consistency.

## 3.6 | Open Database Connectivity

Open Database Connectivity (ODBC), is an application programming interface (API) standard, that makes it possible for applications to access data from a variety of database management systems (DBMS) [7].

ODBC simplifies the process of connecting to databases, executing queries, retrieving data, and managing transactions. It provides a standardized set of functions that work consistently across various platforms and programming languages. This standardization allows developers to write code that can be easily ported across different database systems.

To utilize ODBC, an application requires an ODBC driver. The driver acts as a mediator between the application and the specific DBMS being used, such as OracleDB, MySQL, or Amazon RDS. It handles the translation of ODBC function calls into commands that the DBMS understands, ensuring smooth communication between the application and the underlying database.

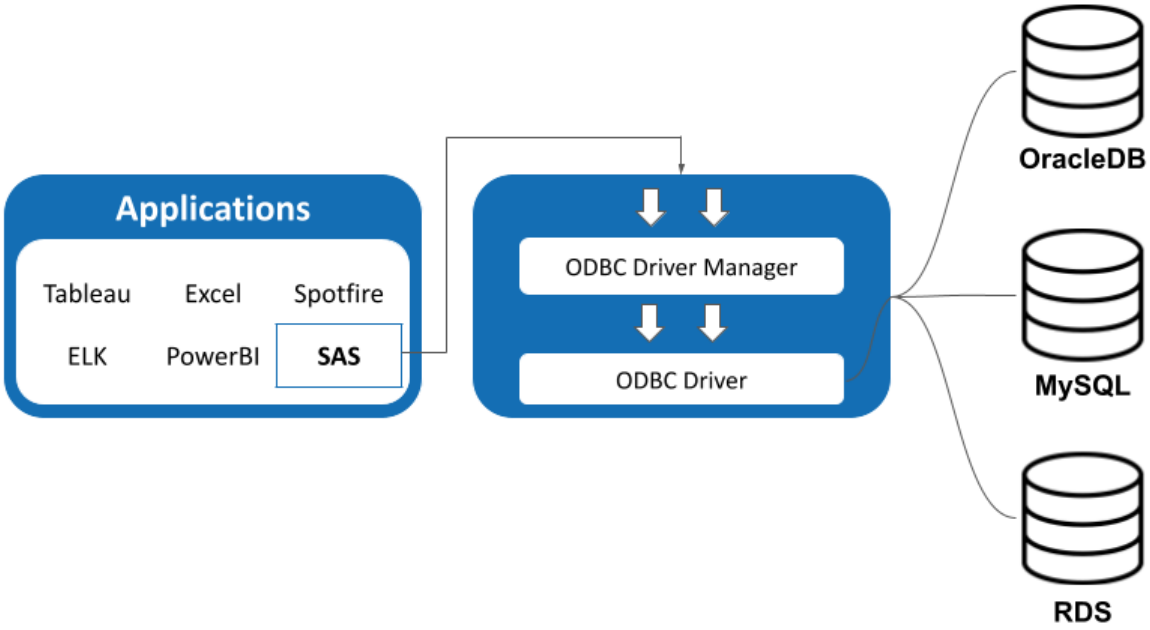


**Figure 3.10:** Open Database Connectivity Pipeline

To simplify, ODBC is a translator that lets the application and database communicate. In terms of SAS, the SAS Drivers for ODBC connect from an ODBC-compliant application to SAS/SHARE servers, SAS Scalable Performance Data (SPD) Servers, or a local instance of a SAS ODBC server.

## 3.7 | Data Quality

Data quality is the conformance of data to the business definitions and the business rules (i.e., business metadata).

### 3.7.1 | Data as a True Asset

Effective planning and management of data is essential, as it is a valuable and reusable asset that can increase in value through integration and analysis. Managing the quality of both data and metadata is crucial, requiring input from cross-functional teams with diverse skills and expertise.

Data possesses unique properties, and its value can be expressed both qualitatively and quantitatively. Taking an enterprise-wide perspective is necessary to fully leverage the potential of data and maximize its value.

### 3.7.2 | Organizational Reasons for Poor Data Quality

The lack of data stewardship and data governance, along with poor metadata management and absence of data lineage, contribute to misconceptions about data quality.

Key data elements may not be clearly defined, and the emergence of new data sources outpaces effective management. Manual data entry without quality checks further hampers data integrity. Insufficient monitoring, reporting, and absence of defined service level agreements for essential business data quality exacerbate the issue. Moreover, the absence of clear data ownership and a culture that recognizes data as a valuable business asset perpetuate these data quality misconceptions.

### 3.7.3 | Data Quality Scope

- Business Definition Quality (*Context*)
  - Business Definitions. Business Rules. Valid Content (valid values, range). Intended Business Purpose (correct usage context). Current Data Quality.
- Data Record Quality (*Rows*)
  - Item completion. Duplication. Cross-table validation. Accuracy.
- Data Element Quality (*Columns*)
  - Domain Integrity: Data type of field (ex. Date). Correct Contextual Value: Business metadata and business rule validation. Accuracy: Source system or expected calculation validation. Cross-Field Validation: Validates the business rule that relates two or more columns or across tables. Format Consistency: Correct format (ex. MM/DD/YYYY).
- Data Movement Quality (*Navigation*)
  - External data quality (incoming – outgoing). Source System to Source System (file transfer, SOA). Source System to data mart / data warehouse. Data Warehouse to Data Marts. To desktops via self-service data. Internal to Cloud – Cloud to Internal.

## 3.8 | Data Governance Journey

### 3.8.1 | Plan: Identify organization needs, opportunities, and efforts

During the planning phase, objectives of the program must be defined and frameworks must be set for how and when decisions will be made.

The planning phase must discuss the initial scope of data governance, objectives, guiding principles, organizational frameworks, roles, responsibilities, and the program charter.

### 3.8.2 | Design: Who makes the decisions and how

During the design phase, decision-making bodies will determine operating procedures and how compliance and progress will be measured. Design methodologies should be decided by an official council of data stewards.

A defined subset of rules should be established that are non-negotiable and must be strictly adhered to. The success of these rules lies in their ability to facilitate the progress of data stewardship, thereby generating tangible value. To validate the effectiveness of your plan, it is recommended to develop a program measurement plan.

This plan involves comparing the current state with previous meetings, using metrics such as steward attendance. For instance, evaluating decision-making based on a required quorum, such as 8 out of 10 members, ensures a robust decision-making process. Furthermore, it is important to note that proxies are not permissible, and the punctuality of members attending the meetings should also be monitored as part of the measurement plan.

A standardized solution to define key activities and assign decision making rights is through a RACI chart.

| Data Governance Function | Activities | PM | DGSC | DGC | DO | OC | DS | DM |
|---|---|---|---|---|---|---|---|---|
| Strategy & Alignment | Set strategic direction and goals for Data Governance | C | A | R | R | C | I | I |
| | Review performance of objectives, goals, status and benefits of the program | A | I | C | R | R | C | I |
| | Coordinate and facilitate Data Governance program | A | I | C | R | R | C | I |
| | Provide program funding | C | A | R | C | C | I | C |
| | Provide overall Data Governance program sponsorship | C | A | R | R | I | I | I |
| | Define Data Governance priorities | C | C | R | A/R | I | I | I |

**Figure 3.11:** A Sample RACI Chart
*R* = Responsible, *A* = Accountable, *C* = Consulted, *I* = Informed

### 3.8.3 | Launch: Kick off initial program – begin small and expand

During the launch phase, the designed operating model will be executed.

1. Onboard participants: Engage and onboard relevant stakeholders, ensuring their understanding of their roles and responsibilities within the project.
2. Execute operating model: Implement the established operating model, which defines the framework and processes for data governance.
3. Policy development and approval: Develop and finalize data governance policies, ensuring alignment with organizational objectives and obtaining necessary approvals.
4. Benefit measurement: Establish a measurement plan to assess the effectiveness of the data governance. This involves defining metrics and tracking progress against predefined goals.

## 3.9 | Policy

A policy refers to a documented set of principles, guidelines, and rules that govern the management, access, usage, and protection of data within an organization.

| Principle | Description |
|---|---|
| Policy | Formal set of statements that define how data resources will be used or managed. |
| Procedure | Detailed instructions about how a policy is to be implemented. |
| Standard | Required configuration that is considered best practice. |
| Best Practice | Technique, method, process, or activity that is more effective at delivering a particular outcome than any other technique, method, process, etc. |
| Data Management | Tactical execution and enforcement of data governance policies and standards. |

**Figure 3.12:** Principle Comparison Against Policy

### 3.9.1 | Policy Statement - Sample

The Data Governance Office, in partnership with the appropriate organization entities and support from IT, will proactively assess, monitor, report, and improve data quality. For the purposes of this policy, data quality is defined as the conformance of data to the definitions and the organization rules (metadata).

## 3.10 | Extract Transform Load (ETL)

Extract-Transform-Load involves three key steps: extracting data from various sources, transforming it to meet specific requirements or standards, and loading it into a target system or data warehouse for analysis, reporting, or other purposes [10].

The ETL process is crucial for ensuring data quality, consistency, and reliability. It enables organizations to integrate data from various sources, standardize it, and make it ready for analysis or reporting purposes.

SAS Data Management Advanced (SAS DMA) is a software suite that provides a comprehensive set of tools and capabilities to support the ETL process, including data integration, data quality management, data governance, and metadata management (5.1).

1. Extract:
   - Data is gathered from multiple sources such as databases, files, APIs, or external systems.
2. Transform:
   - Data is transformation process to ensure consistency, quality, and compatibility with the target system.
3. Load:
   - Data is loaded into the target system or data warehouse.

## 3.11 | Additional: Questions & Answers

To ensure the conversation remains as relevant as possible to UHTASI, dedicated time slots were allocated for question and answer sessions. (Key: *UHTS = UHTASI, SAS = SAS*)

1. UHTS → SAS: *Is there a formal data strategy that outlines, access, integration, movement, storage?*
   - *The inherent issue is that not everything is harmonized across organizations, therefore, there is no unified organizational strategy that solves the problem. Instead, the one responsible for providing a solution that fits the organizational needs is the principle investigator.*
2. SAS → UHTS: *How does UHTASI handle data validation?*
   - UHTASI mainly uses hash check-sums in R and python for data validation. We hope that SAS software alleviate these, or lack-there-of, validation issues. UHTASI checks for missing cells and duplicates in data, however, these cases do not entirely indicate missing data values. Data sources are not standardized amongst our data providers. Therefore, ETL pipelines, data integration, and data quality are also not standardized.
3. SAS → UHTS: *What are common problems that UHTASI's analysts are facing regarding data?*
   - UHTASI's ETL process is dependent on the type of data source that is being used. In the case of federal government data, UHTASI will be provided with an encrypted drive with public and private key authentication processes. However, even a straight-forward solution like the one above requires the tracking and auditing of actions throughout the ETL process.
   - In addition, UHTASI has difficulty getting access to timely data. Data is provided by vendors every quarter, and general trust is established with each vendor.

# 4 | VMware

To create a multi-tenant SAS environment, it is important to familiarize ourselves with the technology utilized by UHTASI in their existing on-premises infrastructure. UHTASI relies on VMware, a company that offers virtualization and cloud computing software solutions. Their primary software, vSphere, is used to build and manage UHTASI's on-premises infrastructure [24].

## 4.1 | 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 [23]. 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:

- **ESXi**: The bare metal hypervisor installed on your machines.
- **vSphere Web Client**: A web-based management interface.
- **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.

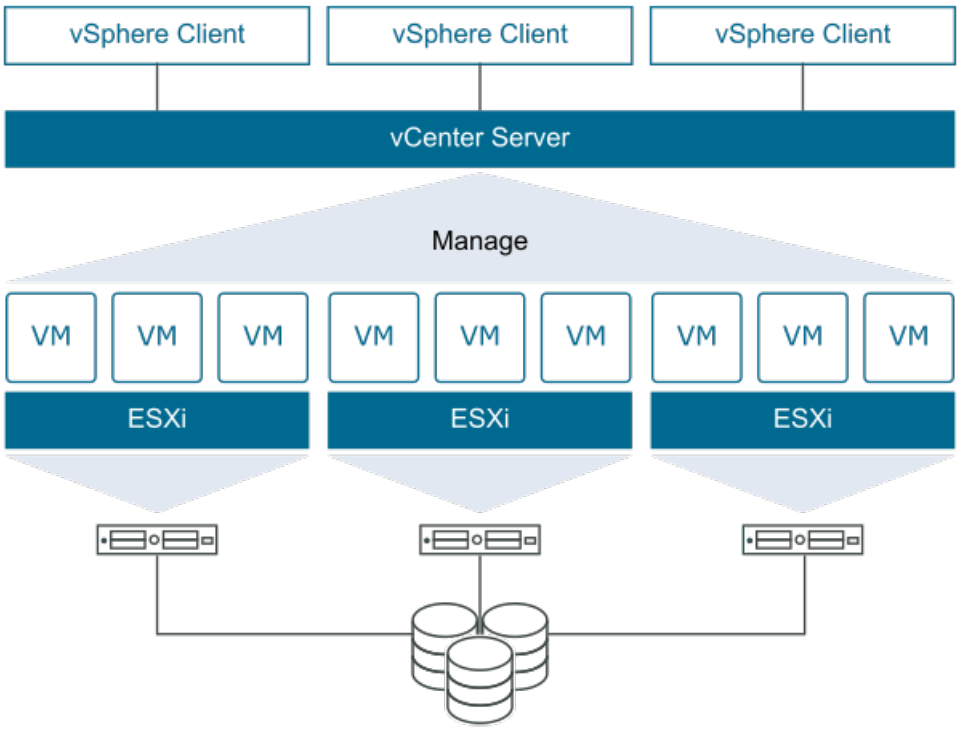


**Figure 4.1:** vSphere's Sotftware Suite Relationship

## 4.2 | ESXi

ESXi is a type-1, bare-metal hypervisor that is installed directly on a server and functions as the primary operating system.

Unlike traditional operating systems (e.g., Linux, Windows Server), ESXi focuses solely on allowing virtualization. While traditional operating systems require the installation of a separate software-based type-2 hypervisor for virtualization, ESXi integrates virtualization directly into the operating system itself.

It's important to note that while ESXi enables virtualization at the OS level, the management of virtualization itself is facilitated through VMware's vSphere suite. vSphere provides the tools and features necessary to create, manage, and optimize virtualized environments. It serves as the management layer for the virtualization capabilities enabled by ESXi.

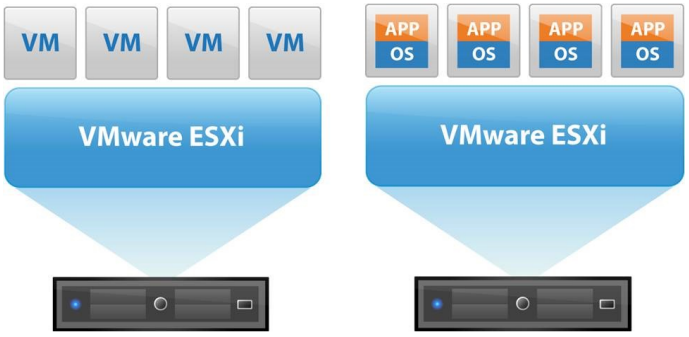


**Figure 4.2:** ESXi

## 4.3 | vSphere Client

The vSphere Client is an application (interface) that allows you to manage and monitor your VMware environments. It is important to note that the actual management capabilities stem from the vCenter Server and **NOT** from vSphere client.

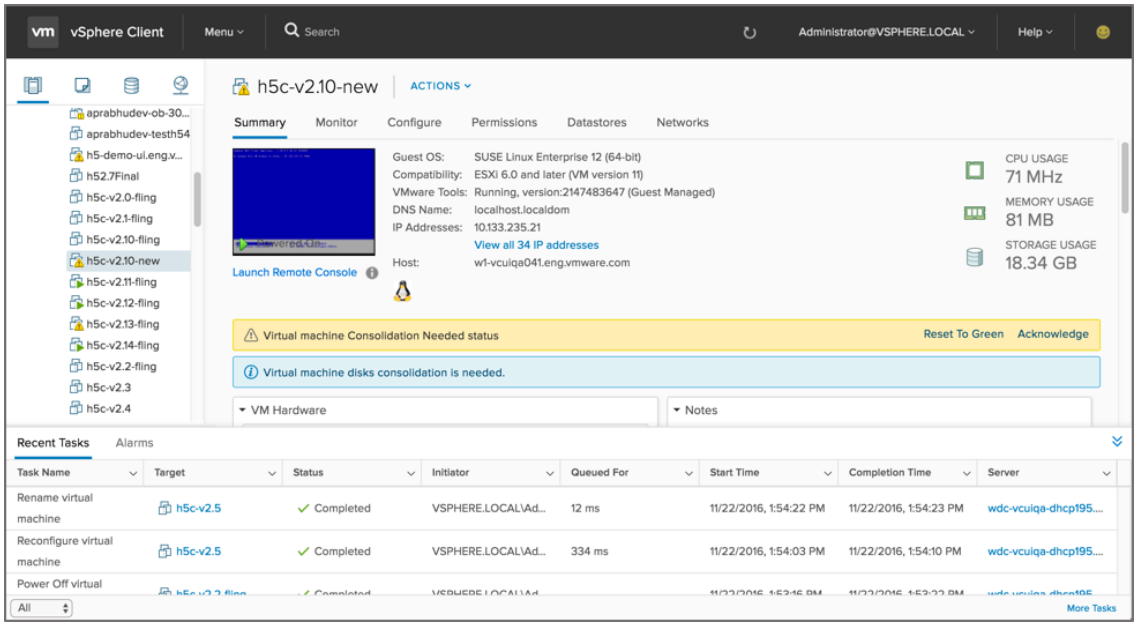


**Figure 4.3:** vSphere Client

## 4.4 | vCenter

The vCenter Server is a centralized platform for managing vSphere environments.

While vSphere serves as the foundation for virtualization, vCenter Server extends these capabilities by acting as a centralized platform designed to manage vSphere environments. Beyond the basic management features offered by vSphere, vCenter Server offers advanced functionalities such as automation, orchestration, and policy-based management.

Some key features of vCenter include: VM migration, security groups, role policies, single sign-on, workflow automation, monitoring and auditing reports, distributed resource management, and optimized resource allocation.

### 4.4.1 | 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[1], auditing[2], encryption[3], secure communication[4], integration[5], and two-factor authentication[6]. These security features help to ensure confidentiality, integrity, and availability of the virtualized infrastructure, a requirement when working with PHI data.

## 4.5 | 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 [25]. 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.

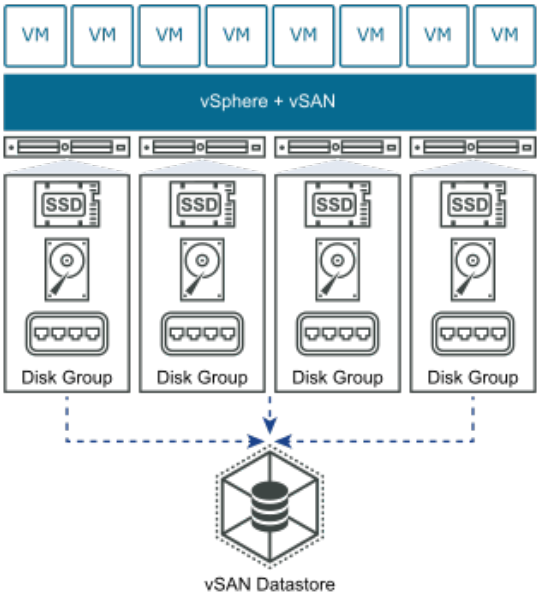


**Figure 4.4:** Standard vSAN Cluster

## 4.6 | 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

[1] Define roles and permissions to users based on their roles to prevent unauthorized access.
[2] Track user activity and changes to identify security issues and log actions taken within the virtualized environment.
[3] Encrypt VM data, configuration files, and communication between hosts.
[4] Supports SSL/TLS encryption to secure communication between hosts and the vCenter server.
[5] Integrate with third-party security products (e.g., antivirus, IDS) to provide additional layers of security
[6] 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.

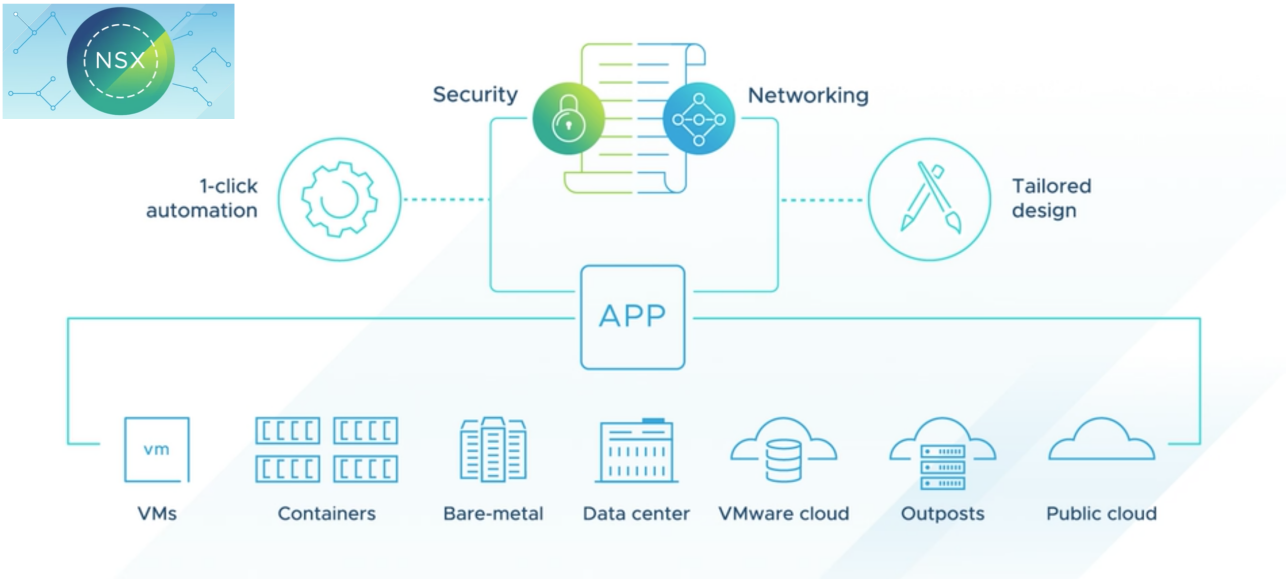


**Figure 4.5:** NSX Infrastructure

## 4.7 | VMotion

VMotion is virtualization software that migrates VMs between physical servers or hosts without disrupting service [14]. The process involves copying the entire state of the VM, including memory, CPU state, and network connections, from one host to another.

To ensure HIPAA compliance when migrating VMs with protected health information (PHI), it is crucial to employ secure protocols and encryption mechanisms to ensure confidentiality, integrity, and availability of PHI during the migration process. IT administrators should utilize encrypted connections or VPN tunnels when transmitting data between the source and destination hosts. By implementing these security measures, host servers and network connections used for VMotion are protected from unauthorized access, mitigating potential compliance issues.

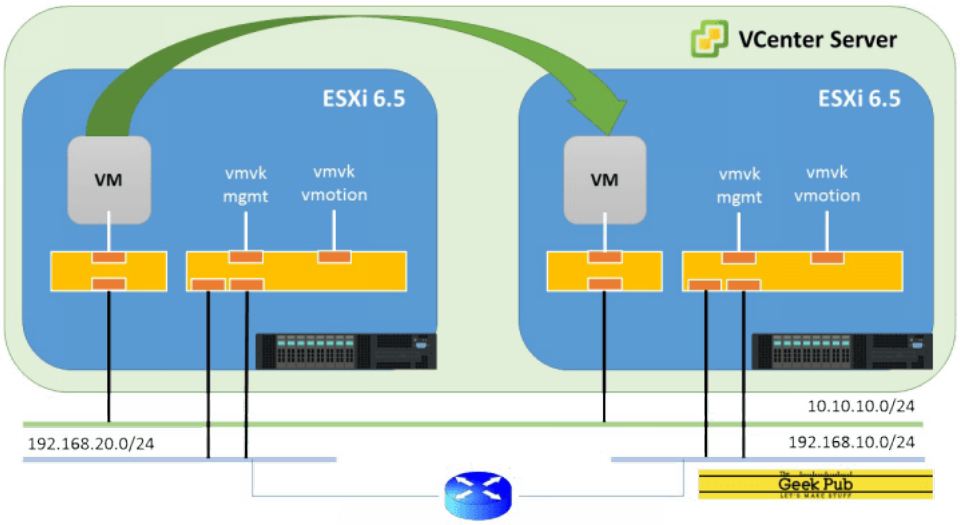


**Figure 4.6:** VMotion

# 5 | Statistical Analysis System (SAS)

Statistical Analysis System (SAS), is a software suite for data quality, data management, and data analytics. In addition, SAS is also a programming language that is designed to be used with SAS technologies to enable data analytics.

UHTASI will be utilizing a range of SAS products suites, and analytic engines to enable data management and data analytics. This includes SAS Data Management Advanced (DMA) as the ETL (3.10) solution, SAS 9.4 and SAS Visual Analytics as the runtime environments, CAS (Cloud Analytic Service) as the analytic engine, and other related SAS products that are embedded in the SAS ecosystem.

## 5.1 | SAS Data Management Advanced (SAS DMA)

SAS Data Management Advanced (SAS DMA), is a software suite that provides a set of tools for data integration and data quality. The purpose of SAS DMA is to support data ETL (3.10) (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 enabled through these three components:

- 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.

## 5.2 | SAS 9.4

SAS 9.4 is a software suite that provides tools for data management, statistical analysis, business intelligence, and predictive modeling [17]. 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.

The following software components will be integrated alongside our SAS 9.4 software suites:

- Base SAS: The foundation of the SAS system and provides essential data management, analytics, and programming capabilities. It includes a powerful programming language, data manipulation tools, statistical procedures, and reporting features [15].
- Data Management Advanced Server: This component encompasses several tools for advanced data management tasks
  - Data Governance - Lineage tracking and workflow management to ensure data governance and compliance.
  - Data Quality - Data standardization, identification of duplicate records, and error detection using visual flow charts and pre-built algorithms. It aims to ensure data quality throughout the entire ETL (Extract, Transform, Load) process and can integrate smoothly with other software.
  - Data Integration Studio - A visual interface for high-level data access, integration, and management tasks, allowing users to create visual flow charts that are converted into SAS code. It is primarily used by IT and administrators.

- Enterprise Guide: This component is designed for low to mid-level data management and access. It offers a visual interface for creating flow charts that are converted into SAS code. It is intended for use by a wide range of users, providing a user-friendly approach to data management tasks.
- SAS/ACCESS to MS: A component of the SAS software suite. It allows users to seamlessly access and interact with Microsoft data sources, such as Excel, Access, SQL Server, and other ODBC (Open Database Connectivity) compliant databases. This integration facilitates data extraction, transformation, and analysis using SAS tools and techniques, enhancing the interoperability between SAS and Microsoft environments.

## 5.3 | 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 [18]. It allows organizations to store, manage, analyze, and share large volumes of data across different sources and formats.

The following software components will be integrated alongside our SAS Viya software suites:

- 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.
- Viya Platform: SAS Viya is a cloud-native and open analytics platform that provides scalable and distributed processing capabilities. It allows users to perform advanced analytics, machine learning, and data visualization tasks in a collaborative and scalable environment.
- Visual Text Analytics: A tool for analyzing unstructured text data, extracting insights, and uncovering patterns and sentiments within text documents.
- Model Manager: A component that facilitates model development, deployment, and monitoring. It enables organizations to efficiently manage and govern their analytical models throughout their lifecycle.
- SAS Optimization: A module that offers optimization techniques to solve complex business problems involving resource allocation, scheduling, logistics, and more. It helps organizations optimize their operations and make data-driven decisions.
- SAS IML (Interactive Matrix Language): A programming language within SAS that provides a flexible and interactive environment for matrix computations, statistical analysis, and modeling. It enables users to perform advanced analytics and create custom statistical models.
- SAS QC (Quality Control): A set of procedures and tools for statistical quality control and process improvement. It helps organizations monitor and analyze data to ensure quality standards and identify areas for improvement.
- SAS Econometrics: A package that provides econometric modeling and analysis capabilities. It allows users to analyze economic data, estimate models, and make predictions or forecasts based on economic relationships.
- Data Preparation: SAS offers various tools and techniques for data preparation, data cleaning, and transformation. These tools help users cleanse and structure their data for analysis, ensuring data accuracy and integrity.

- SAS/ACCESS Engines: SAS provides access engines that enable users to interact with data stored in different database management systems (DBMS). These engines allow SAS to read, write, and manipulate data in various formats and databases.
- Visual Analytics Add-In For Office: An integration that allows users to access and analyze SAS Visual Analytics reports directly within Microsoft Office applications, such as Excel and PowerPoint. It enables users to leverage the power of SAS analytics while working in their familiar Office environment.

To carry out these processes, SAS utilizes the Cloud Analytic Services (CAS) framework.

## 5.4 | 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.

### 5.4.1 | 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).

### 5.4.2 | 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.

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.

### 5.4.3 | 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)**.

### 5.4.4 | Symmetric Multiprocessing (SMP)

The Symmetric multiprocessing (SMP) architecture is used when you want to run CAS on a machine or VM(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 consist of a single VM that serves as both the controller and the 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.

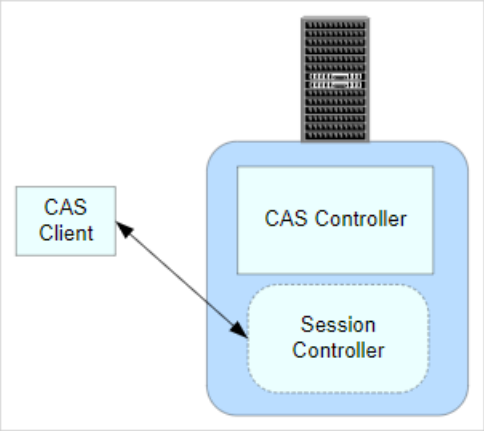


**Figure 5.1:** Single-machine CAS Server

### 5.4.5 | 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.

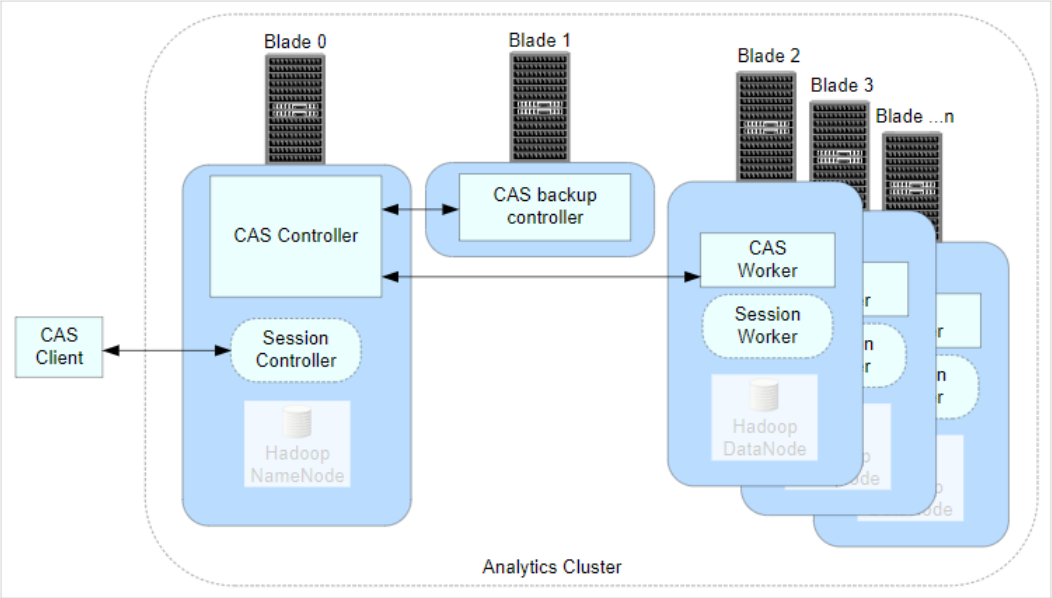


**Figure 5.2:** Distributed CAS Server

## 5.5 | Algorithms & Training

SAS technologies offer a powerful framework for data analytics in various domains, including healthcare. With the availability of pre-built algorithms, provided by esteemed institutions such as the Centers for Disease Control and Prevention (CDC) and the Agency for Healthcare Research and Quality (AHRQ), combined with free training resources, SAS empowers organizations to effectively analyze data and derive meaningful insights, ultimately contributing to evidence-based decision-making and improved healthcare outcomes.

### 5.5.1 | AHRQ & Algorithms

The Agency for Healthcare Research and Quality (AHRQ) plays a vital role in advancing healthcare research and quality improvement initiatives. In conjunction with SAS, AHRQ offers organizations a range of pre-built algorithms or functions, developed by AHRQ's own experts, within the SAS programming language.

These pre-built algorithms apply data analytics or conduct data quality checks on various data sources. By utilizing these algorithms, organizations can streamline their analytic processes and ensure the accuracy and reliability of their data, leading to enhanced research outcomes and more effective healthcare interventions.

- SAS QI v2022 [1]

### 5.5.2 | CDC & SAS Training

The CDC offers comprehensive training and instruction on the SAS programming language, specifically tailored to CDC data, and this training is provided free of charge. By leveraging SAS technologies, organizations can acquire the necessary skills to effectively analyze and interpret CDC data, ultimately contributing to improved public health outcomes and evidence-based decision-making.

- SASSI [5]

### 5.5.3 | CMS & Packages

The Centers for Medicare & Medicaid Services (CMS) is responsible for administering essential healthcare programs in the United States. In collaboration with SAS, CMS provides standard libraries and packages specifically designed for data quality checks within medical-related data sources. By leveraging SAS technologies in conjunction with CMS's libraries, organizations can implement robust data quality checks to ensure the integrity and security of medical-related data.

- Statistical Analysis System (SAS) Package [6]

# 6 | Massively Learning Activities I - Initial Deployment

Although project development plans can differ, they commonly follow a similar framework known as the SDLC (System Development Life Cycle). 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, research, design, implementation, testing & integration, and maintenance. It provides a systematic approach to system development that helps ensure that system is built efficiently with minimal risk.

We explore the logistics of configuring a multi-tenant environment by documenting the entire project management process, inspired by the System Development Life-Cycle framework. Massively Learning Activities will follow a similar variation of the SDLC project management model where each SDLC stage will correspond to a subsection in this chapter.

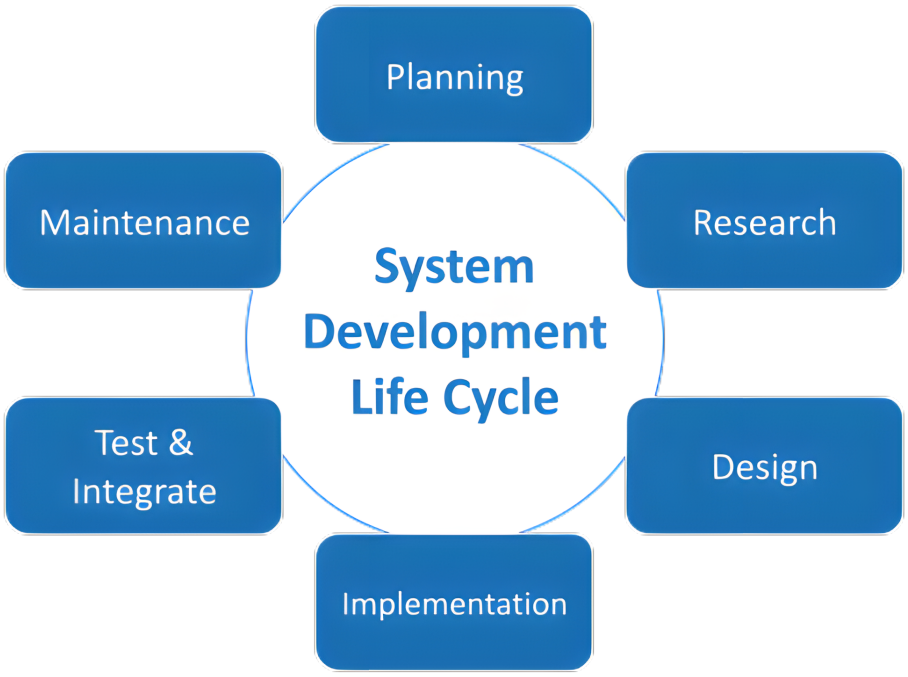


**Figure 6.1:** System Development Life-Cycle

## 6.1 | 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

## 6.2 | 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.

### 6.2.1 | 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. The ITS data center ensures a high level of reliability with a guaranteed 99.8% availability of network, power, and cooling components as stated in the service level agreement (SLA).

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 on the SAN to increase the overall capacity of the SAN's storage. The connection between the network and SAN switches and the Dell FX2 Enclosure is a 10-gigabit link.

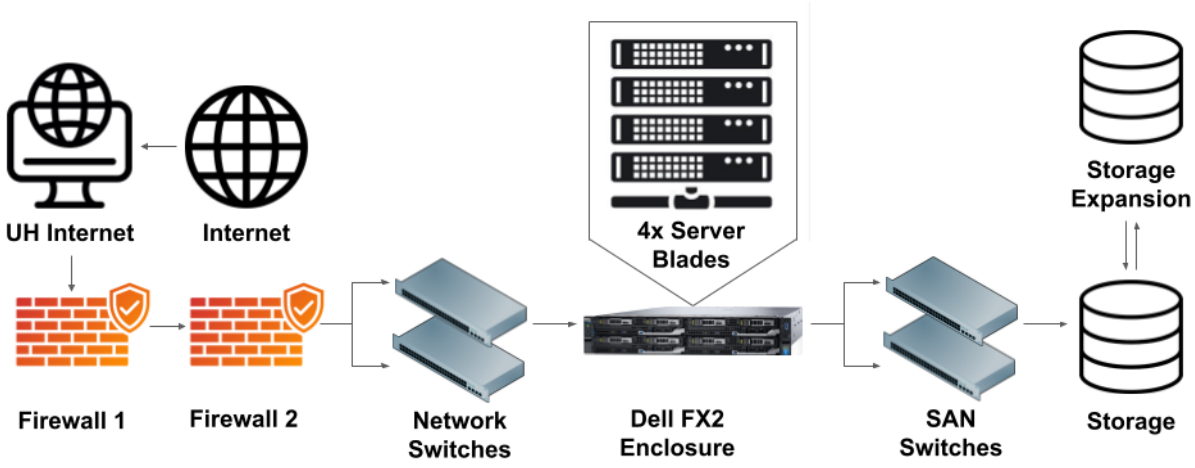


**Figure 6.2:** UHTASI On-Premise Infrastructure

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.

### 6.2.2 | Multi-Tenancy Configuration Plan: VM Location

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

**Figure 6.3:** Physical and logical locations of VMs related to SAS technologies.

### 6.2.3 | Multi-Tenancy Configuration Plan: Resource Allocation

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

**Figure 6.4:** Resource requirements of VMs related to SAS technologies.

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

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

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

### 6.2.5 | 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<br>(8 GB RAM per core minimum) |
| SAS Compute Server | 6 to 8 cores with 48 to 64GB RAM<br>(8 GB RAM per core minimum) |
| SAS Mid-Tier Server (Web-App Server) | 2 cores with 24GB RAM<br>(24 GB RAM per server minimum) |

**Figure 6.5:** 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:
   - As SAS is input/output (I/O) intensive, it is crucial to ensure that the storage environment can meet the required level of I/O throughput as storage is separate from the server, and multiple compute tiers may need access to a common data area.
   - 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.

### 6.2.6 | 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)<br>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<br>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 6.6:** 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.

### 6.2.7 | Pre-Installation Requirements Document: File Path

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

- PATH: \\ AHI-File-Share \ ..300 SAS Installation \ SAS Viya 3.5 \ PIRD \ UHTASI SAS Viya 3.5 PIRD

### 6.2.8 | 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 6.7:** 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 6.3 and 6.4 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<br>· RHEL 8.2 - 8.x<br>· Oracle Linux not supported |
| Operating System Packages and Compliance | · lilbXp & libXmu<br>· numactl package & X11/Xmotif (GUI)<br>· glibc-2.17-107.el7 |
| Key Deployment Information | · Deployed using RPM packages via Ansible config tool<br>· Needs to connect to an LDAP server for authentication |
| Server(s) | · All servers should be dedicated to SAS Viya<br>· All servers should have identical CPU and Memory<br>· Needs to connect to an LDAP server for authentication |
| CPU Guidelines | · Intel Xeon Chipset @ 2.6GHZ<br>· Minimum: 4 cores<br>· Recommended: 16 cores |
| Memory Guidelines | · 16GB of RAM per core @ 1600MHz<br>· 64-96GB of RAM per machine |

**Figure 6.8:** General System Requirements I

| General System Requirements | Metadata |
|---|---|
| Special Considerations for CSP's | · To consult with a SAS sizing expert:<br>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)<br>· Bind: (1) anonymously, (2) with a specific binding account<br>· Recommended to exclude account from password change<br>· Recommended to exclude account from account locking policies |
| Network Considerations | · Recommended 10GB ethernet connection<br>· Servers should have static hostnames and static IP addresses |
| DNS and DNS Alias | · Names need to be resolvable by all hosts in SAS<br>· 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<br>· 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<br>· Ports 80, 443, and possibly 5570, 17551, should be open<br>· Other ports need to be opened for data sources<br>· Refer to Deployment Guide for the complete list |

**Figure 6.9:** General System Requirements II

5. Viya File System:
   - Refer to the Pre-Installation Document for information about the file system (6.2.7).
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 6.9 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 of the following 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<br>· SSH to all Viya hosts from Ansible CTRLR<br>· UID and GID on all SAS Viya Hosts<br>· Either local or domain account |
| SAS Viya Installation User | Yes: sas | · Primary group sas<br>· SSH to all Viya hosts from Ansible CTRLR<br>· Non-expiring password policy<br>· Recommended local user |
| SAS Viya CAS Owner | Yes: cas | · Primary group sas<br>· SSH to all Viya hosts from Ansible CTRLR<br>· Non-expiring password policy<br>· Recommended local user |
| SAS Viya LDAP Bind Account | No (e.g., viya_bind_ldap) | · MUST be an LDAP or Domain Account<br>· Non-expiring password policy<br>· 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 6.10:** SAS Viya 3.5 Service Account Requirements

| Account/Group | Required Name | Required Characteristics |
|---|---|---|
| Tenant Admin Account | No (e.g., acmeadmin, etc) | · Primary group sas<br>· UID and GID on all SAS Viya Hosts<br>· No password assigned<br>· Non-expiring password policy<br>· Recommended domain group |
| Tenant Admin Group | No (e.g., acmeadmgroup) | · Can either be local or domain group<br>· 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<br>· Recommended domain group |

**Figure 6.11:** 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 6.12:** 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.

## 6.3 | 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.

### 6.3.1 | Design Principles

TASI will design a multi-tenant infrastructure that will accommodate all the VMs specified in the multi-tenancy configuration plan in Section 6.2.2. 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**: An excellent application is characterized by agility, reversibility, continuous procedure refinement, proactive failure testing, and organizational learning.
   - UHTASI 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.
   - UHTASI must constantly refine operation procedures, setting up regular game days to validate that all procedures are effective and efficient.
   - UHTASI must anticipate for failure by testing failure scenarios and response procedures of servers, VMs, and SAS components.
   - UHTASI 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.
   - UHTASI'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 UHTASI 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.
   - UHTASI must consider the principle of least privilege when designing an LDAP directory to support multi-tenancy.
   - UHTASI must ensure a robust security framework through Data Governance methodologies such as data classification, access control, privacy compliance, data retention, data quality, data integrity and auditing.

- 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.
   - UHTASI 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, UHTASI must consider load balancing applications for user traffic across multiple servers.
   - To prepare for MLA II (VM Migration), UHTASI must ensure that the initial deployment is loosely 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**:
   - UHTASI will not achieve true cost optimization as MLA I will utilize existing infrastructure for SAS services, while MLA II will involve the procurement of additional hardware through capital expenditures (CAPEX). The hardware utilized in MLA I is expected to reach its end-of-life (EOL) in five years (2028).

### 6.3.2 | IAM Design

**TBD**

### 6.3.3 | 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.

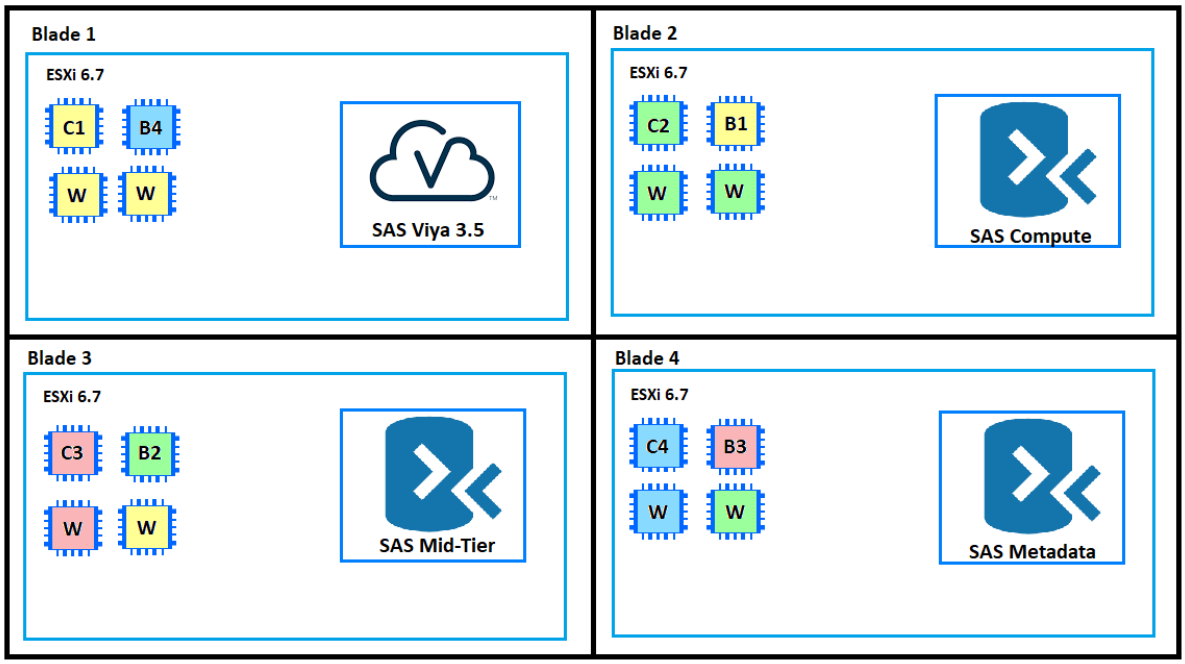


**Figure 6.13:** Multi-Tenant Deployment (**Temporary Figure**)

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.

## 6.4 | Implementation I

The Implementation phase is a pivotal stage in deploying a robust SAS infrastructure following the SDLC framework. This phase involves executing the deployment plan, installing the selected hardware and software components, and configuring the system according to the defined technical specifications. Thorough testing and verification procedures are conducted to ensure the system functions as intended.

| Software | Installed | Date | Status |
|---|---|---|---|
| SAS 9.4 | Completed | April 18, 2023 | Online |
| SAS Data Management Advanced Server | Completed | June 15, 2023 | Online |
| SAS Enterprise Guide | Completed | June 15, 2023 | Online |
| SAS/ACCESS to MS | Completed | June 15, 2023 | Online |

**Figure 6.14:** SAS 9.4 Installation Status

| Software | Installed | Date | Status |
|---|---|---|---|
| SAS Visual Analytics | Pending | TBD, 2023 | Offline |
| SAS Visual Statistics | Pending | TBD, 2023 | Offline |
| SAS Visual Data Mining and Machine Learning | Pending | TBD, 2023 | Offline |
| SAS Visual Forecasting | Pending | TBD, 2023 | Offline |
| SAS In-Memory Statistics | Pending | TBD, 2023 | Offline |
| Viya Platform | Pending | TBD, 2023 | Offline |
| Visual Text Analytics | Pending | TBD, 2023 | Offline |
| Model Manager | Pending | TBD, 2023 | Offline |
| SAS Optimization | Pending | TBD, 2023 | Offline |
| SAS IML (Interactive Matrix Language) | Pending | TBD, 2023 | Offline |
| SAS QC (Quality Control) | Pending | TBD, 2023 | Offline |
| SAS Econometrics | Pending | TBD, 2023 | Offline |
| Data Preparation | Pending | TBD, 2023 | Offline |
| SAS/ACCESS Engines | Pending | TBD, 2023 | Offline |
| Visual Analytics Add-In For Office | Pending | TBD, 2023 | Offline |

**Figure 6.15:** SAS Viya 3.5 Installation Status

## 6.5 | Testing & Integration I

During the Testing and Integration phase, comprehensive testing is performed to validate the functionality, performance, and reliability of the SAS infrastructure. Integration testing is carried out to ensure seamless interaction between different system components and to verify data integrity and accuracy. This phase also includes conducting user acceptance testing to ensure that the system meets the requirements and expectations of end-users.

To ensure the security of UHTASI's system and network, a comprehensive assessment will be conducted through a red team exercise, encompassing both vulnerability scanning and penetration testing. UHTASI will engage a trusted third-party organization (TTPO) to perform the assessment. This approach allows for an unbiased evaluation of UHTASI's security measures and helps identify any vulnerabilities that could be exploited by malicious actors.

### 6.5.1 | Rules of Engagement

Define the scope, objectives, and limitations the TTPO must abide by during the assessment of UHTASI's system and network. As UHTASI handles sensitive PHI, it is crucial to adhere to these rules of engagement to ensure compliance with HIPAA and to avoid damage to data sources (Items listed below are sample rules and not actual limitations):

- Avoid causing any damage or disruption to production systems.
- Exclude specific systems, networks, or assets from testing.
- Limit the scope to a specific set of vulnerabilities or attack vectors.
- Avoid testing third-party systems without proper authorization.
- Prohibit the use of Denial-of-Service (DoS) attacks or any actions that may impact system availability.
- Restrict testing to a specific environment or network segment.
- Exclude social engineering or physical penetration testing.
- Do not attempt to exploit known critical vulnerabilities without prior approval.
- Avoid any activities that violate local or international laws.
- Do not conduct the testing on live customer or user data.

### 6.5.2 | Reconnaissance

**TBD**

### 6.5.3 | Scanning

**TBD**

### 6.5.4 | Vulnerability Assessment

**TBD**

### 6.5.5 | Exploitation

**TBD**

### 6.5.6 | Analysis

**TBD**

### 6.5.7 | Remediation

**TBD**

## 6.6 | Operations & Maintenance I

In the Operations and Maintenance phase, ongoing operational activities and system maintenance tasks are performed to ensure the smooth operation and longevity of the SAS infrastructure. This includes monitoring the system's performance, troubleshooting and resolving any issues, applying necessary updates and patches, and conducting regular backups and data management. Proactive maintenance and periodic system audits are conducted to optimize system performance, enhance security, and adhere to compliance standards.

# 7 | 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 [3].

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[7]/OPEX[8] costs.

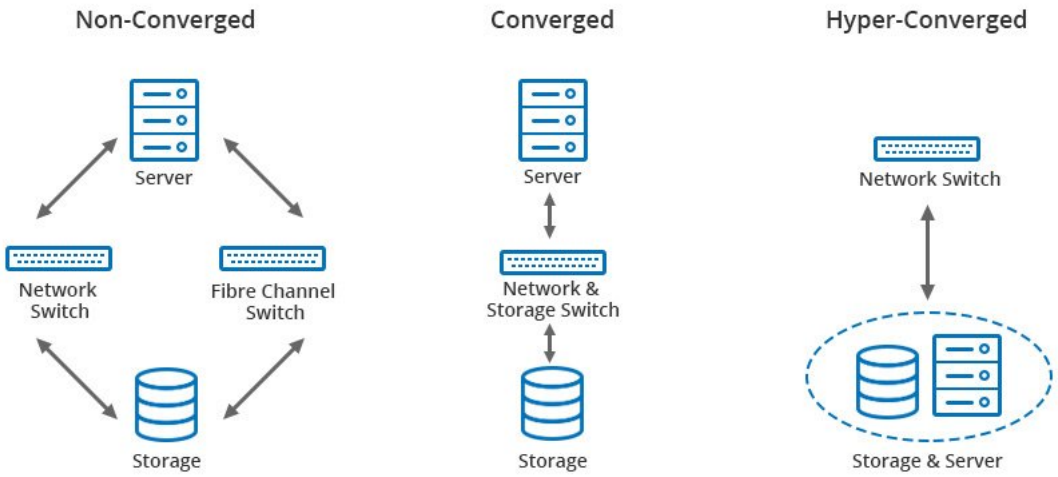


**Figure 7.1:** Types of IT Infrastructures

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.

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

# 8 | Massively Learning Activities II - Migration 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, research, design, implementation, testing & integration, and maintenance. It provides a systematic approach to system development that helps ensure that system is built efficiently with minimal risk.

We explore the logistics of configuring a multi-tenant environment by documenting the entire project management process, inspired by the System Development Life-Cycle framework. Massively Learning Activities will follow a similar variation of the SDLC project management model where each SDLC stage will correspond to a subsection in this chapter.

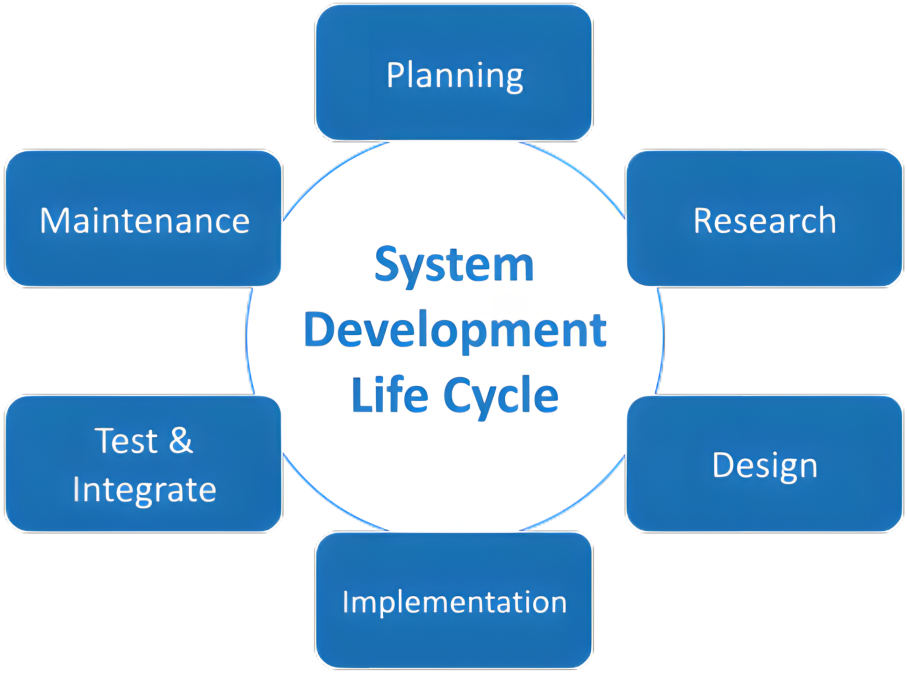


**Figure 8.1:** System Development Life-Cycle

## 8.1 | Planning II

UHTASI has been contracted by CNMI to create an infrastructure that will allow for data analytics on Protected Health Information (PHI). To achieve this, UHTASI 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.

MLA II will primarily focus on the migration of the infrastructure deployed from MLA I, which includes the hosted SAS services and multi-tenancy configuration, to newly acquired hardware in preparation for additional data sources and tenants.

MLA II will expect 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

## 8.2 | Required of Analysis II

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.

**TBD**

## 8.3 | Design II

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.

**The migration process of MLA I to the newly acquired hardware will be completed utilizing VMotion, an VM migration tool provided by VMware.**

## 8.4 | Implementation II

The Implementation phase is a pivotal stage in deploying a robust SAS infrastructure following the SDLC framework. This phase involves executing the deployment plan, installing the selected hardware and software components, and configuring the system according to the defined technical specifications. Thorough testing and verification procedures are conducted to ensure the system functions as intended.

**TBD**

## 8.5 | Testing & Integration II

During the Testing and Integration phase, comprehensive testing is performed to validate the functionality, performance, and reliability of the SAS infrastructure. Integration testing is carried out to ensure seamless interaction between different system components and to verify data integrity and accuracy. This phase also includes conducting user acceptance testing to ensure that the system meets the requirements and expectations of end-users.

To ensure the security of UHTASI's system and network, a comprehensive assessment will be conducted through a red team exercise, encompassing both vulnerability scanning and penetration testing. UHTASI will engage a trusted third-party organization (TTPO) to perform the assessment. This approach allows for an unbiased evaluation of UHTASI's security measures and helps identify any vulnerabilities that could be exploited by malicious actors.

### 8.5.1 | Rules of Engagement

Define the scope, objectives, and limitations the TTPO must abide by during the assessment of UHTASI's system and network. As UHTASI handles sensitive PHI, it is crucial to adhere to these rules of engagement to ensure compliance with HIPAA and to avoid damage to data sources (Items listed below are sample rules and not actual limitations):

- Avoid causing any damage or disruption to production systems.
- Exclude specific systems, networks, or assets from testing.
- Limit the scope to a specific set of vulnerabilities or attack vectors.
- Avoid testing third-party systems without proper authorization.
- Prohibit the use of Denial-of-Service (DoS) attacks or any actions that may impact system availability.

- Restrict testing to a specific environment or network segment.
- Exclude social engineering or physical penetration testing.
- Do not attempt to exploit known critical vulnerabilities without prior approval.
- Avoid any activities that violate local or international laws.
- Do not conduct the testing on live customer or user data.

### 8.5.2 | Reconnaissance

**TBD**

### 8.5.3 | Scanning

**TBD**

### 8.5.4 | Vulnerability Assessment

**TBD**

### 8.5.5 | Exploitation

**TBD**

### 8.5.6 | Analysis

**TBD**

### 8.5.7 | Remediation

**TBD**

## 8.6 | Operations & Maintenance II

In the Operations and Maintenance phase, ongoing operational activities and system maintenance tasks are performed to ensure the smooth operation and longevity of the SAS infrastructure. This includes monitoring the system's performance, troubleshooting and resolving any issues, applying necessary updates and patches, and conducting regular backups and data management. Proactive maintenance and periodic system audits are conducted to optimize system performance, enhance security, and adhere to compliance standards.

**TBD**

# 9 | References


[1] AHRQ. AHRQ - Quality Indicators. Qualityindicators.ahrq.gov, 2022.

[2] AWS. AWS Identity & Access Management. Amazon Web Services, Inc., 2017.

[3] Shaikh Azeem and Satyendra Sharma. Study of converged infrastructure & hyper converge infrastructre as future of data centre. *International Journal of Advanced Computer Research*, 8:900, 09 2019.

[4] Billmath. What Is Identity Lifecycle Management with Azure Active Directory? - Microsoft Entra. Learn.microsoft.com, 15 Mar. 2023.

[5] CDC. Self-Directed and STD-Focused SAS Instruction (SASSI). www.cdc.gov, 5 Aug. 2021.

[6] CMS. QualityNet Home. Qualitynet.cms.gov, 2022.

[7] David-Engel. What Is ODBC? - ODBC API Reference. Learn.microsoft.com, 28 Feb. 2023.

[8] Jai Narayan Goel and B.M. Mehtre. Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology. *Procedia Computer Science*, 57:710–715, 2015.

[9] Google. What Is Data Governance? Google Cloud.

[10] IBM. ETL (Extract, Transform, Load). www.ibm.com, 2022.

[11] IBM. What Is Encryption? Data Encryption Defined. www.ibm.com, 2022.

[12] Vijay Khatri and Carol V. Brown. Designing Data Governance. *Communications of the ACM*, 53(1):148, 2010.

[13] Microsoft Security. What Is Identity Access Management (IAM)? — Microsoft Security. www.microsoft.com.

[14] Mike Murray. How vmotion works! (vmotion explained). The Geek Pub, April 2017.

[15] SAS. Base SAS. Support.sas.com.

[16] SAS. Programming Documentation for the SAS® Viya® Platform. Go.documentation.sas.com, 1 June 2023.

[17] SAS. Programming Documentation for SAS® 9.4 and SAS® Viya® 3.5. Documentation.sas.com, 31 Mar. 2021.

[18] SAS. Introduction to SAS Cloud Analytic Services. Documentation.sas.com, 5 Apr. 2023.

[19] Secret Double Octopus. What Is LDAP & Active Directory? How LDAP Works — Security Wiki. doubleoctopus.com.

[20] TASI/PHIDC. TASI/PHIDC. Social Science Research Institute.

[21] Daniel Teachey. Data Governance Framework: What Is It and Do I Already Have One? Sas.com, 2019.

[22] U.S. Department of Health & Human Services. HIPAA for Professionals. HHS.gov, 2021.

[23] VMware. What's new in vsphere 6.5: vcenter management clients. VMware vSphere Blog, December 2016. Accessed on 27th June 2023.

[24] VMware. VMware vSphere Documentation. Vmware.com, 2019.

[25] VMware. vsan deployment options. Vmware.com, 2023.

# A | Appendix A: Acronyms & Abbreviations Glossary

| Acronym | Meaning |
|---|---|
| ABAC | Attribute-Based Access Control |
| APCD | All-Payer Claims Database |
| CAS | Cloud Analytic Services |
| CAPEX | Capital Expenditures |
| CFS | Clustered File System |
| 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 |
| DoS | Denial Of Service |
| EHR | Electronic Health Record |
| EOL | End Of Life |
| 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 |
| ITS | Information Technology Services |
| I/O | Input / Output |
| LDAP | Lightweight Directory Access Protocol |
| MPP | Massively Parallel Processing |
| OPEX | Operational Expenditures |
| 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 |
| SLA | Service Level Agreement |
| SMA | State Medicaid Agency |

| Acronym | Meaning |
|---|---|
| SMP | Symmetric Multi-Processing |
| SSL | Secure Sockets Layer |
| SSO | Single Sign-On |
| SSRI | Social Science Research Institute |
| TASI | Telecommunications and Social Informatics Program |
| TBD | To Be Determined |
| TLS | Transport Layer Security |
| TTPO | Trusted Third Party Provider |
| UH | University of Hawaii |
| VM | Virtual Machine |