\Pontificia Universidad Católica del Ecuador

Facultad De Ingeniería

Escuela de Sistemas

TEMA:

Diseño de infraestructura tecnológica para la gestión y procesamiento de datos biológicos.

AUTOR:

Pérez Vera Michael Alexander

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS DE INFOMRACIÓN

QUITO, 08-09-2022

DEDICATORIA

El presente proyecto de integración curricular se la dedico a mi familia que gracias a su apoyo pude concluir mi carrera.

A mis padres y hermanas por su apoyo y confianza. Gracias por ayudarme a cumplir mis objetivos como persona y estudiantes. A mi padre por brindarme los recursos necesarios y enseñarme como enfrentar las adversidades y dificultades. A mi madre por hacer de mí una mejor persona a través de sus consejos, enseñanzas y amor. A mis hermanas por estar siempre presentes, acompañándome para poderme realizar. A mi pareja por darme el tiempo para realizarme profesionalmente.

AGRADECIMIENTO

A mi familia, por darme todo su apoyo y quererme por sobre todas las cosas

A Ayliz por darme su amor, apoyo y confianza y compartir nuevos e inolvidables momentos en mi vida, te quiero mucho.

A Oscar por siempre estar a mi lado apoyándome y brindarme su amistad sincera.

A mi tutor Charles Escobar sin usted y sus virtudes, su paciencia y constancia este trabajo no lo hubiere logrado tan fácil.

RESUMEN

El presente proyecto de integración curricular realiza un diseño de infraestructura para la gestión y procesamiento de datos biológicos mediante la investigación y comparativa de las diferentes tecnologías de procesamiento, almacenamiento y comunicación. Así mismo se hace énfasis en las diferentes formas de despliegue de arquitectura donde se planea demostrar las ventajas que se tiene en cada una de ellas.

Estos indicadores fueron empleados y comparados para mostrar relación y diferencia entre distintas formas de evaluar sus características y desempeño en el manejo de datos masivos, puesto que estos indicadores permiten observar la diferencia existente entre ellas y facilitan en la toma de decisiones.

En el primer capítulo se describe la justificación del trabajo, se plantea la problemática existente, se establece el objetivo general y los objetivos específicos y el alcance de esta. En el capítulo 4 se demuestra el cumplimiento de los objetivos planteados al principio. Terminando con conclusiones y recomendaciones para la mejor toma de decisión de la infraestructura.

ÍNDICE

Tabla de contenido

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS

ÍNDICE DE TABLAS

CAPÍTULO I: INTRODUCCIÓN

1. MARCO DE REFERENCIA

1.1. JUSTIFICACIÓN

El proyecto BIOINCA maneja una cantidad de datos masiva debido a las secuencias de ADN. El uso de esa información representa una oportunidad sobre la cual buscan realizar una gestión y procesamiento de datos para obtener información útil y veraz en el menor tiempo posible, de tal forma que cualquier investigación que se realice se enfoque más tiempo en la comprensión y uso en vez del procesamiento de los datos a ser estudiados. Bajo este paradigma, surge la necesidad del proyecto en identificar una infraestructura IT que sea capaz de integrar, y procesar esa data con el fin de acelerar el proceso de preparación de los datos correspondientes a los casos de estudio que se desarrollen. La factibilidad de implementar una infraestructura de esa magnitud es el primer paso para la implementación en las instalaciones de la universidad, lo que implica realizar un estudio, comprobar los requisitos mínimos que deben disponer los componentes que conformaran el ambiente de trabajo para el estudio de las secuencias de ADN pueda brindar resultados útiles y en el menor tiempo disponible para los investigadores, independientemente del caso de estudio que se ejecute en el momento.

El estudio de cadenas de ADN en el proyecto BIOINCA abarca un conjunto de información en la cual se engloba datos correspondientes a las cadenas de ADN de las especies estudiadas, de este estudio con múltiples ejemplos y casos se presenta la necesidad de identificar el ambiente tecnológico adecuado para el estudio y tratamiento de la información recopilada de determinadas especies. Como parte de la búsqueda donde se abarca una gran variedad de información, se plantea elaborar un estudio en el cual se presente un estudio de los requerimientos mínimos y necesarios para el procesamiento de datos biológicos, para ello

se estudia las necesidades físicas de Hardware que se necesitaría en la Pontificia Universidad Católica del Ecuador, con el fin de determinar si dentro de las características presentadas por los equipos, la información pueda ser gestionada, controlada, manejada y procesada de tal forma que su estudio represente una oportunidad de visualización y procesamiento de datos biológicos para uso de los integrantes del proyecto BIOINCA, de tal manera que se pueda solicitar a la Universidad la infraestructura que permita realizar estudios y en tiempo real, coordinados bajo las necesidades que cada grupo interno presenten en su momento.

Por lo tanto, se necesitará realizar un estudio de factibilidad de infraestructura IT que ayude en la gestión y procesamiento de datos biológicos, para ello será necesario identificar los servicios que los biólogos realizan en la gestión y procesamiento de la data. Para poder cumplir con esto se necesitará determinar la arquitectura de procesamiento y almacenamiento adecuada; tanto el procesador como núcleo(s) principal(es) que realizarán los cálculos respectivos y la memoria RAM necesaria para transferir y permitir procesar esa información.

Adicionalmente se debe determinar con que opciones se contaría para la unidad de almacenamiento principal o disco duro del cual se requiere determinar cuál la mejor tecnología para el almacenamiento y/o recuperación de la información.

A su vez toda esta información tiene que ser transmitida por algún medio de comunicación donde se debe analizar la mejor solución de rendimiento en conectividad de red para que la carga y descarga de información sea estable y los tiempos menoren.

Dado que se cuenta con varios posibles despliegues de infraestructura se hace necesario comparar la infraestructura tradicional, contenedores y la virtualización en busca de una un equilibrio que considere costos, capacidad de procesamiento y almacenamiento.

1.2. Planteamiento del problema

Debido a que las secuencias de ADN que se estudian en el proyecto BIOINCA generan una cantidad masiva de información se ha visto la necesidad de una infraestructura IT necesaria para la gestión y procesamiento de esa información, sin embargo, no se cuenta con ella y tampoco se ha realizado un estudio de factibilidad de esta.

Dado que activamente se realizan investigaciones se genera una gran cantidad de datos a ser procesados, los cuales aumentan diariamente. A esto se suma el hecho de que en caso de que se quiera implementar esta infraestructura, no se han levantado las necesidades de servicios para las actividades que realizarían los integrantes del proyecto, por lo que no se ha podido encontrar soluciones de infraestructura que correspondan a las necesidades que tienen, por lo tanto para desarrollar sus actividades han tenido que buscar soluciones en otras instituciones que cuenten con la infraestructura necesaria para el procesamiento de la información, aunque al no contar con una arquitectura de procesamiento, almacenamiento, comunicación y gestión, su uso puede tener un costo para la Universidad y los biólogos. Adicionalmente al no contar con un estudio comparativo sobre las alternativas de despliegue entre las tecnologías disponibles como la virtualización, contenedores y la infraestructura tradicional, se hace complicado para la Universidad una implementación dado que no cuenta con un estudio de factibilidad que ayude a la misma.

De lo discutido, se identifica el siguiente problema principal:

- No se cuenta con un estudio de factibilidad de infraestructura IT para la gestión y procesamiento de datos biológicos.

Y los siguientes problemas secundarios:

- No se conoce la arquitectura que se podría usar para el procesamiento y gestión de datos biológicos.

- No se ha identificado la infraestructura de almacenamiento para gestión y procesamiento de datos biológicos.
- No se cuenta con una infraestructura de comunicación para transferencia de datos biológicos.
- No se cuenta con una comparativa de tecnologías de virtualización vs contenedores e infraestructura tradicional.

1.3. Objetivo General

Realizar un estudio de factibilidad de infraestructura IT para la gestión y procesamiento de datos biológicos para el proyecto BIOINCA.

1.4. Objetivos Específicos

1.4.1. **Identificar los servicios usados por biólogos para la gestión y procesamiento de datos biológicos**

1.4.2. **Identificar la arquitectura de procesamiento para la gestión y procesamiento de datos biológicos.**

1.4.3. **Identificar la infraestructura de almacenamiento para gestión y procesamiento de datos biológicos.**

1.4.4. **Identificar la infraestructura de comunicación para gestión y procesamiento de datos biológicos.**

1.4.5. **Comparar las tecnologías de virtualización vs contenedores.**

1.5. Antecedentes

“El Laboratorio Internacional en Biodiversidad y Agricultura Sustentable en los Andes Tropicales (BIOINCA) es una plataforma de investigación internacional, inter-institucional e interdisciplinaria patrocinada por el Instituto Nacional Francés para el Desarrollo Sustentable

(IRD), la Universidad de Los Andes (Uniandes-Colombia) y la Pontificia Universidad Católica del Ecuador.” (PUCE INVESTIGA, s. f.)

“Acoge a un equipo de 50 científicos colombianos, ecuatorianos y franceses de sólida experticia en temas sobre genomas, socio-ecosistemas y el estudio del cultivo natural y artificial de biosistemas de plantas en los Andes tropicales. Su objetivo de la investigación está organizado en el estudio de: a) Las características que definen a la biodiversidad, tanto cultivos alterados por el hombre, como naturales; b) el estudio de las interacciones biológicas, c) el estudio de ecosistemas, servicios y la búsqueda de soluciones naturales; y, por último, d) el desarrollo de herramientas y programas de entrenamiento innovadores, ecológicos y bioinformáticos.” (PUCE INVESTIGA, s. f.)

1.6. Alcance

Este proyecto está orientado a levantar las necesidades del equipo de investigación BIOINCA, donde se presentará diseño de infraestructura tecnológica y las características que necesitaran los componentes para que pueda soportar la gestión de procesamiento, almacenamiento y comunicación de datos biológicos, a su vez se presentará la comparativa entre las distintas opciones de despliegue para la infraestructura.

2. Marco Teórico

La criticidad de los datos afecta a toda la infraestructura de TI; esto debería ser una consideración importante al planificar su estrategia de gestión de datos. Es importante asegurarse de que ningún conjunto de datos individual tenga un poder descontrolado sobre el resto del ecosistema de aplicaciones o TI. (¿Qué es infraestructura de TI?, s. f.)

2.1. Procesadores:

"La Unidad Central de Proceso o CPU (Central Processing Unit) es el componente encargado de interpretar las instrucciones de los programas y procesar los datos. También se le conoce como procesador o microprocesador. Es un componente fundamental de un ordenador que ha estado presente desde sus inicios. Pero no es hasta la década de los 70 cuando se fabrican los primeros procesadores a partir de circuitos integrados." (ConceptoABC, s. f.)

2.1.1. Intel x86_64

La arquitectura Intel® 64 ofrece computación de 64 bits en los diseños integrados cuando se combina con el software de soporte. (ConceptoABC, s. f.)

2.1.2. AMD x86_64

AMD64 es una arquitectura de procesador de 64 bits desarrollada por Advanced Micro Devices (AMD) para añadir capacidades informáticas de 64 bits a la arquitectura x86. (AMD64, s. f.)

2.1.3. ARM

Arquitectura que, en principio, se muestra más eficiente que el x86 para fabricar dispositivos móviles, pero tiene la dificultad de conseguir rendimientos elevados por encima de ciertos valores de frecuencia de reloj y voltajes. (ConceptoABC, s. f.)

2.2. Memoria Principal

"(Random Access Memory), Memoria de Acceso Aleatorio.

Es volátil (su contenido se pierde si se interrumpe la corriente eléctrica que la alimenta), permite tanto leer como escribir datos rápidamente a través de señales eléctricas y el acceso a una posición se efectúa de modo aleatorio." (ConceptoABC, s. f.)

Se trata de una memoria en la que puede leer y escribir un equipo o cualquier otro dispositivo. (ConceptoABC, s. f.)

2.2.1. RDIMM

DIMM registrado, un módulo de memoria dual en línea con registros.

2.2.2. UDIMM

DIMM sin búfer, es decir, un módulo de memoria en línea dual sin búfer.

2.2.3. LRDIMM

Load Reduced DIMM, un módulo de memoria dual en línea de baja carga.

2.3. Entrada Salida E/S

"Entrada/salida o E/S, hace referencia a la comunicación entre un sistema de procesamiento de información (como un equipo con Symphony instalado), y el mundo exterior (posiblemente una persona o cualquier otro sistema de procesamiento de información, como un sistema de control de acceso)" (Dispositivo: E/S, s. f.)

2.3.1. Unidad de Almacenamiento

“Una unidad de almacenamiento es un dispositivo de lectura o escritura de datos digitales. Generalmente, éstos son capaces de guardar archivos de manera permanente, aunque también están diseñados para poder eliminar y gestionar los archivos que se almacenan en él.

Las unidades de almacenamiento necesitan de un soporte físico que permita el almacenamiento de información, además, los archivos que se recogen en él deben cumplir con condiciones informáticas que hagan posible su almacenamiento.” (ConceptoABC, s. f.)

2.3.1.1. HDD

Un disco duro (también conocido como disco rígido) es un dispositivo de almacenamiento de datos capaz de contener información digital de distintos formatos. Las unidades de disco duro son capaces de guardar datos aún si están apagados. Esto les convierte en una pieza clave para lograr la funcionalidad de las computadoras y distintos sistemas informáticos. (ConceptoABC, s. f.)

2.3.1.2. SSD

Los discos de estado sólido, SSD por sus siglas en inglés, son dispositivos que almacenan los datos de un ordenador. Su nombre hace referencia al hecho que no cuenta con ninguna movilidad mecánica en su interior y se caracteriza principalmente por no tener platos ni discos magnéticos, como los discos duros convencionales (HDD). Además, aporta a los ordenadores mayor velocidad y latencia de la que podrían ofrecer los HDD. (Glosario informático - Definición de términos informáticos, s. f.)

2.3.1.3. NVME

Memoria no volátil express o NVMe (por sus siglas en inglés, Non-Volatile Memory Express), es un protocolo de almacenamiento, desarrollado recientemente para discos de estado sólidos (SSD) sirviéndose del puerto PCI-Express. Esto representa una de sus principales

características, ya que logra una mayor velocidad en transferencia de datos, por tener un mayor ancho de banda. (Glosario informático - Definición de términos informáticos, s. f.)

2.4. Comunicación computacional

“La comunicación es el intercambio de datos entre computadoras a través de una conexión. Para que las computadoras puedan entenderse debe haber un "lenguaje" común llamados protocolos.” (ConceptoABC, s. f.)

2.4.1. Ancho de banda

“En sistemas digitales, el ancho de banda digital es la cantidad de datos que pueden ser transportados por algún medio en un determinado período de tiempo (generalmente segundos). Por lo tanto, a mayor ancho de banda, mayor transferencia de datos por unidad de tiempo (mayor velocidad)” (Glosario informático - Definición de términos informáticos, s. f.)

2.4.2. SSH

SSH File Transfer Protocol, es un protocolo de red que provee acceso, administración y transferencia de archivos sobre un flujo de datos fiable (especialmente SSH). (Glosario informático - Definición de términos informáticos, s. f.)

2.4.3. Conexión remota

el acceso remoto es acceder desde una computadora a un recurso ubicado físicamente en otra computadora, a través de una red local o externa (como internet). (Glosario informático - Definición de términos informáticos, s. f.)

2.5. Alternativas del Despliegue de Infraestructura

Intro

2.5.1. Virtualización

Es una simulación: ejecuta software en equipos donde no está instalado, usa procesadores que no están presentes y guarda archivos en espacios "virtuales".

La virtualización es un enfoque para implementar recursos de computación que aíslen las diferentes capas; hardware, software, datos, redes, almacenamiento; unas de otras. (Glosario informático - Definición de términos informáticos, s. f.)

2.5.1.1. On-Premises

En las instalaciones. En el centro de datos de la empresa. En las instalaciones generalmente se refiere a la realización de una operación en casa en lugar de en instalaciones de terceros, como un proveedor de la nube. (PCMag, s. f.)

2.5.1.2. Nube

“Computación en la nube. Significa desarrollo basado en Internet y el uso de la tecnología de computación.

Se trata de servicios que ofrecen las funcionalidades de programas que antes se tenía en el ordenador en servidores remotos a través de Internet.

Es un estilo de computación donde las capacidades relacionadas con IT se proveen "como un servicio", permitiendo a los usuarios acceder a servicios habilitados tecnológicamente "en la nube" sin conocimiento de, o experiencia en, o control sobre la infraestructura tecnológica que los soporta.” (PCMag, s. f.)

2.5.2. Contenedores

Una arquitectura de servidor que permite que varias aplicaciones y servicios se ejecuten en sus propias particiones aisladas. Los contenedores tienen menos sobrecarga que la infraestructura común de máquinas virtuales, y las aplicaciones se lanzan más rápidamente. (PCMag, s. f.)

CAPÍTULO III: METODOLOGÍA

Para el presente trabajo de titulación se utilizará artículos científicos, tesis, libros, revistas, la cual servirá como referencias bibliográficas, el método que se aplicará es una implementación hibrida entre la recolección de información cuantitativa y cualitativa debido a que es la mejor solución para obtener datos de las necesidades tecnológicas que presentan los biólogos del proyecto BIOINCA con el fin de poder presentar una propuesta de diseño de infraestructura.

El diseño se desarrollará usando datos comparativos en cada uno de los campos a estudiar tales como procesamiento, almacenamiento, comunicación y despliegue, los mismos que devolverán un resultado en base a los parámetros escogidos para cada uno de ellos.

CAPÍTULO IV: DISEÑO DE INFRAESTRUCTURA TECNOLÓGICA PARA LA GESTIÓN Y PROCESAMIENTO DE DATOS BIOLÓGICOS.

4.1. Levantamiento de servicios usados por biólogos para la gestión y procesamiento de datos biológicos

Para poder determinar los servicios usados por los biólogos, se ha recurrido a la fuente primaria de quienes usarían las herramientas, docentes, estudiantes de grado y postgrado e investigadores que utilizarán la infraestructura, para ello se ha recurrido al levantamiento de información a través de entrevistas y encuestas, que se detallan a continuación:

- Entrevista mediante una serie de preguntas abiertas, las que se prepararon con el fin de conocer y ahondar en las actividades que los biólogos realizan con la información (data) que recolectan al extraer los genes de las distintas especies animales o plantas que extraen para analizarlas y estudiarlas.
- Encuesta mediante una serie de preguntas cerradas, se diseñaron dos encuestas diferentes para los dos grupos, los biólogos y los técnicos, donde se buscó profundizar en los temas puntuales para el cumplimiento de los objetivos.

En base a las encuestas se pudo determinar que las actividades realizadas por los biólogos que corresponden a las ejecuciones computacionales son:

- Extracción de las bases: Corresponden a las cadenas genéticas biológicas, las mismas tienen diferentes fuentes de obtención tales como bases de datos en línea, extracción de especímenes obtenidos localmente, nano puertos o extraídos de terceros. La misma que puede llegar a tener 180 gigas.
- Preprocesamiento de las bases: En base a los objetivos y alcance definido por los biólogos las bases extraídas requieren ser analizadas como flujo de datos en bruto

para lo cual se puede requerir varias herramientas, las más usadas para bases grandes y pequeñas o segmentos son:

- Herramientas de análisis de datos en BASH

  Las mismas que están enfocadas en bases biológicas grandes, debido a que la memoria RAM limita el máximo de tamaño asignable a una variable con una gran cantidad de información, aunque ayuda a trabajar con mayor fluidez con datos grandes igualmente existen limitaciones de hardware.

- Herramientas de análisis de datos en R

  Las herramientas desarrolladas en este lenguaje han sido optimizadas para trabajar con bases pequeñas o segmentos de bases grades, las mismas que han sido desarrolladas para ejecutar en menor tiempo y con mayor precisión las técnicas de preprocesamiento.

- Herramientas de análisis de datos en Python

  Las herramientas desarrolladas en este lenguaje dinámico, popular y multifacético han permitido que sea posible trabajar con bases pequeñas o segmentos de bases grades y dependiendo de las optimizaciones de la herramienta con bases grandes, pudiendo realizar las ejecuciones en tiempos reducidos sin perder la precisión en las técnicas de preprocesamiento.

- Herramientas de análisis de datos de terceros

  Desarrolladas y optimizadas para trabajar con bases grandes, diseñadas para utilizar toda la capacidad computacional del equipo y siendo eficaces con la información masiva que recibe. A pesar de que puede ser usada en bases de menor tamaño, se obtiene una eficacia mejor en un entorno para el cual fue elaborado.

- Procesamiento de las bases: Una vez preparada la data se requiere extraer la información en base a los criterios de mapeo, búsqueda, reducción y comparación.

- Mapeos: Visualización de las cadenas de ADN o proteínas.
- Búsqueda: Búsqueda de un argumento en base a toda la data.
- Reducción: Filtrado o extracción de una porción de información.
- Comparación: Comparativa de un argumento en base a toda la data.

Actividad	Definición	Requerimientos	Consideraciones
Extracción de las bases	Corresponden a las cadenas genéticas biológicas, las mismas tienen diferentes fuentes de obtención tales como bases de datos en línea, extracción de especímenes obtenidos localmente, nano puertos o extraídos de terceros.	Almacenamiento Conectividad	• Capacidad • Velocidad R/W • Ancho de banda
Procesamiento de las bases	En base a los objetivos y alcance definido por los biólogos las bases extraídas requieren ser analizadas como flujo de datos en bruto para lo cual se puede requerir varias herramientas, las más usadas para bases grandes y pequeñas o segmentos	R Python BASH Scripts propios Herramientas de terceros	• Sistema Operativo • Versiones • Entornos y dependencias • Contenedores • Replicabilidad

Procesamiento de las bases	Una vez preparada la data se requiere extraer la información en base a los criterios de mapeo, búsqueda, reducción y comparación.	R Python BASH Scripts propios Herramientas de terceros	• Sistema Operativo • Versiones • Entornos y dependencias • Contenedores • Replicabilidad

Tabla 1 Levantamiento de servicios

4.2. Análisis de arquitectura de procesamiento para la gestión y procesamiento de datos biológicos.

Este estudio se ha elaborado a base de entrevistas e información recolectada de diferentes fuentes donde se busca solventar las necesidades tecnológicas de procesamiento de información para los datos biológicos que se han analizado mediante los resultados de las necesidades de los usuarios para el procesamiento de datos biológicos.

En base a estás necesidades se pueden considerar las tres plataformas de arquitectura existentes y comparar entre ellas para determinar lo que mejor se ajuste a las necesidades según los siguientes parámetros:

- Cuota de mercado: El enfoque de calificación es servidores, data centers, supercomputadores.
- Compatibilidad con las herramientas biológicas: Se enfoca en la cantidad de librerías disponibles.
- Costos: Coste por núcleo y coste por Gigahercio.
- Multinúcleo: Varios núcleos en el diseño y su aprovechamiento.
- Eficiencia energética: Coste de gigahercio por vatio consumido.

La escala de calificación se puntúa de 1 a 5 siendo 1 el más bajo y 5 el más alto.

Procesadores	Cuota de Mercado	Compatibilidad con las Herramientas Biológicas	Costos	Multinúcleo	Eficiencia energética	Total
Intel x86_64	3	5	3	5	4	20
AMD64	4	5	3	5	5	22
ARM	2	3	4	5	2	16

Tabla 2 Análisis comparativo entre procesadores

A la vista de este resultado obtenido en la tabla 2 según los parámetros establecidos anteriormente se puede observar que AMD es la arquitectura más sobresaliente de las tres, siendo Intel la segunda mejor por un margen bastante ajustado, por lo que en vista de los datos la arquitectura que mejor se adapta a las necesidades estudiadas es AMD.

Otra de las necesidades latente y complementarias al procesamiento es la RAM, la misma que puede tener varias opciones de arquitectura dependiendo del modelo del servidor, frecuencias aceptadas y modelo de fabricación de las cuales se hará una comparación de los modelos más usados para data center y super computadores para determinar lo que mejor se ajuste a las necesidades según los siguientes parámetros:

- Rendimiento: Mejor manejo del paso de la información.
- Ancho de banda y latencia: Cantidad de información que puede transmitir y el tiempo que le toma hacerlo.
- Costo: Coste por Megahercio.
- Eficiencia energética: Coste de megahercio por vatio consumido.

RAM	Rendimiento	Ancho de banda y latencia	Costo	Eficiencia energética	Total
UDIMM	3	2	2	3	10
RDIMM	4	3	4	3	14
LRDIMM	5	4	5	4	18

Tabla 3 Análisis comparativo entre módulos de RAM

En base a los datos registrados en la tabla se puede determinar dos opciones bastante buenas a considerar los módulos de memoria RAM RDIMM y LRDIMM, tomando en consideración el coste beneficio y el rendimiento se opta que la segunda opción es la más rentable tomando en consideración que el precio es alto por el mismo diseño de su arquitectura que trabaja mucho mejor, aunque se deba disminuir un poco las frecuencias se ganar una mejor optimización de la carga y eficiencia energética.

Factores	On-Premise	Cloud
Alojamiento y gestión de servicios	X	X
Cifrado de datos altamente seguro	X	X
Latencia baja a media	X	X
Rápida actualización de hardware	X	
Flexibilidad y escalabilidad de recursos		X
Alta inversión inicial e inversión regular en infraestructura	X	
Visibilidad y control de datos	X	X
Copias de seguridad y recuperación de datos automatizadas		X
Implementación rápida		X

Tabla 4 Análisis comparativo entre Nube vs On-Premise

Como se puede observar en la comparación entre una solución on-premise y la nube, las dos soluciones están bastante parejas según estos parámetros, considerando que la nube ofrece una ventaja, que es la rápida implementación esto se traduce en reducción de tiempos para un proyecto y aprovechamiento en la etapa de pruebas.

4.3. Comparativa de tecnologías de almacenamiento para gestión y procesamiento de datos biológicos.

Este estudio se ha elaborado a base de entrevistas e información recolectada de diferentes fuentes donde se busca solventar las necesidades tecnológicas de almacenamiento de información para los datos biológicos que se han analizado mediante los resultados de las necesidades de los usuarios para el procesamiento de datos biológicos.

En base a estás necesidades se pueden considerar las tres modelos de almacenamiento existentes y comparar entre ellas para determinar lo que mejor se ajuste a las necesidades según los siguientes parámetros:

- Velocidad: Velocidad máxima de lectura y escritura.
- Temperatura máxima: Temperatura máxima optima de funcionamiento sin reducir el tiempo de vida del disco.
- Consumo energético: Consumo máximo de energía en uso.
- Ruido: Decibeles de ruido producido por el disco en funcionamiento.
- Peso: Peso promedio medido en gramos.

Parámetros	NVME	SDD	HDD SAS
Velocidad	7000 MB/s	650 MB/s	377 MB/s
Vida útil máxima	1200 TBW	1600 TBW	5 años
Temperatura máxima	30 °C	40 °C	60 °C
Consumo energético	0.08 watts	0.279 watts	14 watts
Ruido	0 db	0 db	80 db
Peso	10 gr	40 gr	1102.25 gr

Tabla 5 Análisis comparativo entre tecnologías de almacenamiento

Como se observa en la tabla el disco NVME es la mejor opción por temas de velocidad y consumo energético, pero hay que considerar que la temperatura es un aspecto clave a largo

plazo para reducir el desgaste de los componentes y el tiempo de vida útil del mismo. Por otro lado, analizando que los discos HDD SAS al ser diseñados con ciclos de escritura y lectura más amplios que los otros. Por lo tanto, un aspecto importante a tomar en cuenta es una combinación de los dos tipos de almacenamiento, el NVME para ejecuciones del procesamiento de datos y preprocesamiento y el HDD SAS para un almacenamiento a largo plazo.

4.4. Comparativa de tecnologías de comunicación para gestión y procesamiento de datos biológicos.

Este estudio se ha elaborado a base de entrevistas e información recolectada de diferentes fuentes donde se busca solventar las necesidades tecnológicas de comunicación tanto interna como externa para los datos biológicos que se han analizado mediante los resultados de las necesidades de los usuarios para el procesamiento de datos biológicos.

En base a estás necesidades se pueden considerar las distintas vías de comunicación existentes y comparar entre ellas para determinar lo que mejor se ajuste a las necesidades. Se analizó la mejor vía de enlace para comunicación y el ancho de banda requerido usando el tamaño de archivo registrado en los datos encuestados.

- Acceso remoto

Parámetros	SSH	Conexión remota	Interfaz Web	Jupyter Web
Seguridad	X		X	X
Autentificación	X	X	X	X
Mejor manejo de la privacidad de información	X		X	X
Mayor número de Vulnerabilidades		X		X
Alto consumo de ancho de banda	X	X		

Tabla 6 Análisis comparativo de las vías de conexión

- Ancho de banda

	Tamaño del archivo					
Tamaño del ancho de banda / Tiempo seg	**10 Mb**	**100 Mb**	**500 Mb**	**1 Gb**	**10 Gb**	**100 Gb**
10 mbps	8	80	400	819	8192	80000
100 mbps	0	8	40	81	819	8000
1 gbps	0	0	3	8	80	800
10 gbps	0	0	0	0	8	80

Tabla 7 Análisis comparativo del tiempo de descarga y subida de información

Como se muestran las métricas en la tabla 7, las soluciones de 1 Gbps y 10 Gbps son las más adecuadas para trabajar con la carga y descarga de bases. Tomando en estimación que los servidores se encuentren en América, valores que serán aumentados si se tomamos en consideración el país del que se descargue.

4.5. Comparativa de las tecnologías de virtualización vs contenedores.

Para realizar la comparación en la siguiente tabla se toman como indicadores las necesidades planteadas como el almacenamiento, comunicación y procesamiento de los datos biológicos, a su vez se tomará en cuenta el despliegue, el soporte y mantenimiento de las diferentes tecnologías para abarcar no solo características de rendimiento sino de funcionalidad y eficiencia.

Los Indicadores comparativos:

- Procesamiento
- Comunicación
- Almacenamiento
- Despliegue
- Soporte y Mantenimiento

	VENTAJAS	DESVENTAJAS
VIRTUALIZACIÓN	Bajo coste. Recuperación ante desastres. Menor tiempo de inactividad. Administración centralizada. Optimización de espacio. Escalabilidad. Menor consumo de Energía. Rápido Mantenimiento. Mejor Rendimiento. Tolerancia a Fallos.	Perdida de rendimiento. Requiere más memoria si se requiere varias MV al mismo tiempo. El hardware que no se administre con el hipervisor no puede ser virtualizado. La obsolescencia del hardware virtual. Bajo coste de mantenimiento.

CONTENEDORES	Modularidad. Capas y control de versión de la imagen. Restauración. Rápida implementación. Optimización de espacio. Tolerancia a Fallos. Rápido despliegue Administración centralizada. Menor tiempo de inactividad. Bajo coste. Rápida comunicación entre contenedores.	Gestión compleja a mayor número de contenedores. Complejo manejo de los procesos. Depende de un SO host donde desplegarse.

Tabla 8 Ventajas y desventajas de las diferentes formas de despliegue

	VIRTUALIZACIÓN	CONTENEDORES
Bajo coste	X	X
Recuperación ante desastres	X	X
Menor tiempo de inactividad		X
Administración centralizada	X	X
Soporte y Mantenimiento		X
Optimización de espacio		X
Escalabilidad		X

Menor consumo de Energía		X
Mejor Rendimiento	X	
Tolerancia a Fallos	X	X
Rápido Mantenimiento		X
Restauración		X
Rápido despliegue		X
Intercomunicación		X

Tabla 9 Análisis comparativo entre virtualización vs Contenedores

Cada uno de los indicadores con los que se busca trabajar fueron cubiertos por cada una de las tecnologías mostradas anteriormente, teniendo en cuenta los datos mostrados en la tabla de análisis cuantitativo observamos las fortalezas que los contenedores brindan teniendo este una calificación 13/14 que está muy por encima de la virtualización que obtiene un puntaje de 5/14.

Tecnología / marca	Características	Obsolescencia	Costo
On premise / Intel	HDD 8TB / Xeon Silver 4309Y 2.3 Ghz / 128 GB RAM	4 años	22.500,00 US
On premise/ AMD	HDD 8TB / AMD EPYC 7302P 3.0 Ghz / 128 GB RAM	4 años	20.850,00 US
Nube /	HDD 8TB / 64 vCPUs / 128	Modelo On Demand / Proveedor	23.692,00

Contenedores	Gb Ram	mantiene la infraestructura actualizada	USD / año

Tabla 10 Comparativa de costos de despliegue

Según la tabla 10 se observa los costos representativos entre dos tipos de soluciones y siendo la opción on premiese entrega dos alternativas Intel y Amd. Se observa que los costes de inversión inicial en las dos primeras son bastante altos, por otro lado, la opción en la nube ofrece una solución de pagos mensualizada igualmente bastante alta para las actividades que se realizan dentro del área investigativa del grupo BIOINCA, pero tomando en cuenta que no se incluye gastos de mantenimiento ni soporte, gastos de repuestos y actualización, la opción en la nube es la más rentable analizando costo, rendimiento y beneficio.

CONCLUSIONES

- Durante el estudio para la generación del diseño fue importante identificar y separar todos los servicios que se necesitan para funcionar de manera correcta. De esta forma, se pudo identificar cada uno de los componentes independientes para cumplir con la carga o peticiones que reciba para que no haya afectación para el resto de las ejecuciones.
- Según los datos mostrados para obtener un rendimiento óptimo para trabajar con bases biológicas se debe tener en cuenta los datos mostrados en el capítulo 4 en donde según el estudio realizado se ha determinado cuales son las mejores opciones para escoger para plantear una solución viable para la necesidad presente que tiene el grupo de investigación BIOINCA.
- No siempre la opción con mejores resultados es la opción más viable dado que al tener en cuenta más de un factor, tal como el almacenamiento donde una esquematización combinada entre velocidad y duración se juntan para ofrecer una mejor alternativa que cada una de ellas por separado.
- Uno de los puntos más críticos a tomar en cuenta es la necesidad de contar con un ancho de banda dedicado de por lo menos 11 Gbps dado que la información que se maneja con los archivos puede llegar a ser bastante grande y los tiempos de carga y descarga de las mismas pueden tomar mucho tiempo y llega a variar mucho dependiendo de la zona donde el enlace de descarga se encuentre.
- La Nube es una opción moderna que brinda no solo despliegues más rápidos, sino que a su vez entrega mejores alternativas y menores riesgos a largo plazo, tales como la obsolescencia del hardware, los mantenimientos y los respaldos, siendo esto lo más crítico de este proyecto.

RECOMENDACIONES

- Se debe tomar en consideración que la opción en la nube ofrece una ventaja competitiva tanto en coste como en implementación las mismas que en este momento post pandemia se encuentran en auge y a la orden dado el rápido crecimiento de las empresas nacientes y resuelve las necesidades tecnológicas de las empresas que se adaptan al cambio presente.
- Se recomienda incorporar una combinación hibrida en el almacenamiento para tener eficiencia y disponibilidad a largo plazo dado que al combinar la velocidad de los discos NVME y la larga vida útil de los HDD SAS se tiene las dos necesidades cubiertas.
- Por otro lado, se debe tener en cuenta la necesidad de interfaces de comunicación que soporten 1 Gbps o adoptar por tecnologías de 2.5 Gbps dado el rápido desarrollo de nuevas tecnologías de comunicación que para el fin al que se va a dedicar es bastante necesaria ya que nos ofrece un mejor rendimiento, velocidades más altas y tiempos más cortos. Sin olvidar que no solo el ancho de banda es suficiente, sino que se adopte la conexión por fibra óptica para maximizar aún más esas ventajas.

BIBLIOGRFÍA

- *Alegsa.com.ar - Portal de informática, internet, tecnologias y web*. (s. f.). https://www.alegsa.com.ar
- *AMD64*. (s. f.). SUSE Defines. https://www.suse.com/suse-defines/definition/amd64/
- ConceptoABC. (s. f.). *ConceptoABC - Conceptos de la A a la Z*. https://conceptoabc.com/
- *Dispositivo: E/S*. (s. f.). https://www.aimetis.com/webhelp/Symphony/6.14/es/Dispositivo_E_S.htm
- *Glosario informático - Definición de términos informáticos*. (s. f.). https://www.glosarioit.com/
- PCMag. (s. f.). *Encyclopedia Index*. PCMAG. https://www.pcmag.com/encyclopedia
- ¿Qué es infraestructura de TI? (s. f.). https://www.ibm.com/es-es/topics/infrastructure