Cuba Salud
IV Convención
Internacional de Salud
17-21 de octubre, 2022

Propuesta de diseño de arquitectura de software para la gestión, análisis y procesamiento de datos de neurociencia

Jesús Enrique Fuentes González[1]
Dr.C. Arturo Orellana García[1]

[1] Centro de Informática Médica, Universidad de las Ciencias Informáticas, La Habana, Cuba, contactos: aoreallana@uci.cu, jesusefg@estudiantes.uci.cu

Resumen: El diseño de la arquitectura para la gestión, procesamiento y análisis de neurodatos está concebido para solventar los problemas de interoperabilidad, mantenimiento, escalabilidad y funcionalidad presentes en los softwares que se encuentran funcionando en los centros de investigación de neurociencias. Estos problemas son carecterísticas o métricas intrínsecas en la calidad del desarrollo de software vínculadas a la arquitectura de software y su diseño. La investigación se centra en presentar el diseño de la arquitectura de software así como una explicación de sus componentes con el objetivo de mitigar la carencia de estas métricas de calidad. Para la elaboración del trabajo se siguió una estrategia explicativa y se emplearon los métodos: análisis documental, inductivo-deductivo y modelado del resultado.

La arquitectura de software descrita permite la adquisición automática y manual de datos de neurociencias con diferentes formatos, la automatización de flujos de trabajo y procesos de investigación e innovación, ejecución en entornos remotos de alto rendimiento de múltiples algoritmos y procedimientos computacionales asociados al preprocesamiento de datos y al aprendizaje automático, también permite la visualización y el análisis estadístico de los datos de neurociencia que gestiona. Presenta potencialidades para convertirse en un sistema de diagnóstico asistido por computadoras.

Palabras clave: Arquitectura de software, automatización, datos de neurociencia, diagnóstico asistido por computadoras**.**

I. INTRODUCCIÓN

La neurociencia es el campo de la ciencia que estudia el sistema nervioso (1) y todos sus aspectos, por ejemplo: estructura, función, desarrollo ontogenético y filogenético, bioquímica, farmacología y patología, y como sus elementos interactúan dando lugar a la cognición y la conducta (2–5). A medida que ha cobrado importancia para la comunidad científica la neurociencia se ha dividido en varias ramas de investigación, entre ellas: neurociencia afectiva (6), celular (7), computacional (8), molecular (9), neuroquímica (10), neuroanatomía (11), nanoneurociencia (12) y neuroimagen (13). Estas ramas convergen tangencialmente en el campo conocido como neuroinformática (14), en el cual se involucran ingenieros informáticos y científicos de la computación que orientan sus esfuerzos a la construcción de herramientas para la adquisición de datos, aumentar la capacidad de análisis y construir modelos matemáticos que permitan ordenar, gestionar y agilizar el proceso de experimentación e investigación, impactando significativamente en la mejora del acceso, la eficiencia, la eficacia y calidad de los procesos investigativos.

Este creciente desarrollo tecnológico en los campos de la medicina se evidencia a partir de la construcción de nuevos equipos de cómputo y adquisición de información desarrollado por decenas de fabricantes (Philips, Siemens, General Electric, Kodak y otras) así como técnicas más precisas y especializadas para analizar los estudios generados produciendo un incremento no despreciable en los formatos y tipos de datos obtenidos de los centros médicos y de investigación (15). Al aumento de la heterogeneidad de los formatos de datos en el campo de las neurociencias se le añade el volumen de los mismos y la velocidad requerida para transportarlos por sistemas de cómputo generando problemas de adquisición, almacenamiento y procesamiento asociados comúnmente al *Big Data (16).*

Debido a este incremento de la complejidad en la gestión de datos se han desarrollado plataformas que aglutinan un subconjunto de herramientas y funcionalidades con el fin de satisfacer las necesidades de los procedimientos llevados a cabo por los investigadores y científicos. Plataformas como LORIS (17) y CBRAIN (18) son las disponibles actualmente para ejecutar la gestión y procesamiento de neurodatos en los centros de investigación cubanos, que al tener en cuenta los grandes volúmenes de datos y como consultarlos se han establecido como las plataformas más usadas en Cuba para la gestión parcial de datos de neurociencias y su procesamiento.

Durante el uso de estas herramientas los centros de investigación han generado diferentes procesos de adquisición, depuración y análisis de datos, sus propios formatos para manejar estudios, procedimientos innovadores y nuevas formas de gestionar y compartir el conocimiento. Este proceso de innovación propio de los centros de investigación genera una demanda de escalabilidad, mantenibilidad, reusabilidad, adaptabilidad e interoperabilidad que no cumplen CBRAIN, LORIS ni las herramientas y sistemas de software que se encuentran disponibles, haciendo que se acumule deuda técnica en los sistemas rápidamente, propiciando la obtención de resultados de investigación opacos con costosas consecuencias para los centros de investigación e investigadores entre las cuales destacan: pérdida de datos, extenuantes procesos manuales sujetos a errores e incapacidad para compartir información.

La escalabilidad, mantenibilidad, reusabilidad, adaptabilidad e interoperabilidad son métricas propias de la calidad en el desarrollo de software y están vinculadas a la arquitectura de los mismos (19), lo cual lleva a replantearse el diseño de los software de gestión, análisis y procesamiento de datos de neurociencia. El presente trabajo tiene como objetivo proponer el diseño de una arquitectura de software para la

gestión, procesamiento y análisis de datos de neurociencia teniendo en cuenta las necesidades de los centros de investigación.

II. MÉTODO

Para la ejecución de la presente investigación se sigue una estrategia explicativa y se emplearon los métodos: análisis documental para obtener datos e información asociados al objeto de estudio. Se analizaron los documentos bibliográficos referentes a soluciones arquitectónicas, sistemas de información, su integración, interoperabilidad, así como estándares y buenas prácticas, lo que permitió establecer los fundamentos teóricos. Mediante el método inductivo-deductivo se pudo arribar a conclusiones generales sobre los diseños arquitectónicos propuestos en sistemas homólogos en función de identificar los impactos del sistema en las diferentes aristas propuestas en la investigación. Se aplicó el método de modelado para representar el diseño de la arquitectura de software.

A. Comparación de diferentes plataformas y sistemas homólogos

La propuesta ha sido diseñada a partir del estudio de las plataformas de gestión de datos de neurociencia que tiene como objetivos fundamentales determinar las características que pueden ser reutilizables y recopilar las buenas prácticas y los errores más comunes usados en proyectos ya establecidos en el campo. En (20) se muestra la comparación de varias plataformas de almacenamiento y análisis de datos y no brindan información sobre los diferentes aspectos técnicos de cada una.

Tabla 1: Comparación entre diferentes características de las plataformas de almacenamiento. Fuente adaptado de (20).

	HPC	**Computación de rejilla**	**Computación en la nube**
Arquitectura	Homogénea	Heterogénea	Heterogénea
Organización	Programador de trabajos	Organización Virtual	Máquinas Virtuales o Contenedores
Escalabilidad	Baja	Media	Alta
Red	Dedicada, de alta velocidad	Red de máquinas virtuales	Dedicada, de alta velocidad
Sistemas Operativos	Linux/Unix/BSD	Linux/Unix/Windows	Linux/Unix/Windows
Distribución de los recursos	Centralizado	Distribuido	Híbrido
Almacenamiento de datos	NFS o sistemas de archivos paralelos	Sistemas hechos a medida con GridFs	Soporta varios sistemas de archivos y bases de datos
Administrador de procesos	Los procesos entran en la cola y esperan a ser programados y ejecutados	Buscan los recursos apropiados para ejecutar aplicaciones especialmente grandes	Gestionan los recursos en tiempo real, dependiendo de la demanda
Aplicaciones	Ciencia, educación, negocios y computación empresarial	Ciencia, educación, negocios y computación empresarial	Servicios web, distribución y almacenamiento

También se compararon las métricas de escalabilidad, mantenibilidad, interoperabilidad y reusabilidad de los siguientes sistemas:

- **LORIS** (*Longitudinal Online Research and Imaging System*): es un sistema de gestión de datos modular y extensible basado en la web que integra todos los aspectos de un estudio multicéntrico: desde la adquisición de datos heterogéneos (imágenes, clínicos, conductuales y genéticos) hasta el almacenamiento, procesamiento, y finalmente la difusión (17).

- **CBRAIN** (Canadian Brain Imaging Research Platform, CBRAIN): plataforma que ofrece acceso transparente a fuentes de datos remotas, sitios informáticos distribuidos y una variedad de herramientas de procesamiento y visualización dentro de un entorno seguro y controlado (18).
- **Ecosistema abierto para neurodatos en la nube** o *"NeuroData's Open Data Cloud Ecosystem"* fue una propuesta de Fark y Vogelstein para solventar una serie de problemas asociados al aplicar enfoques tradicionales a la arquitectura de sistema de gestión de neurodatos (19).

Tabla 2: Comparación cualitativa entre diferentes métricas de calidad de software de sistemas homólogos. Fuente: los autores.

	Escalabilidad	**Mantenibilidad**	**Interoperabilidad**	**Reusabilidad**
LORIS	Baja	Baja	Media	Baja
CBRAIN	Alta	Media	Alta	Media
Ecosistema Abierto en la Nube para Neurodatos	Alta	Alta	Alta	Media

Las tablas anteriores muestran que a medida que los sistemas se se orientan a ser distribuidos, con enfoque de microservicios y a la nube, mejoran los estándares de calidad y la adaptabilidad de los mismos. Son capaces de gestionar recursos en tiempo real y soportan varias fuentes de datos, ideales para servicios web, distribución de almacenamiento y procesamiento. LORIS muestra un sistema robusto para la gestión parcial de datos de neurociencia, pero capacidades reducidas para escalar que pueden aumentar al integrarse con CBRAIN. CBRAIN tiene alto rendimiento y capacidades para escalar horizontalmente a partir de la integración con múltiples computadoras de alto rendimiento y el Ecosistema Abierto en la Nube para Neurodatos puede escalar horizontalmente, vertical y por replicación, presenta alta mantenibilidad e interoperabilidad, características propias del estilo arquitectónico orientado a microservicios.

III. RESULTADOS

Teniendo en cuenta los datos anteriores se propone el diseño de una arquitectura de microservicios dirigida por eventos (EDM) para la gestión, procesamiento y análisis de los neurodatos. Dado que el estilo arquitectónico de microservicios nació en la industria aún existen grandes brechas entre el actual nivel, presente en la industria y la academia, por lo que no se han determinado estándares para los estilos de modelado de esta arquitectura (20). Sin embargo, los diagramas son usados con frecuencia en la literatura para representar los microservicios entre los cuales se identifican:

- Diagrama de componentes/contenedores.
- Diagrama de procesos.
- Diagrama de secuencias.
- Diagrama de despliegue.
- Diagrama de clases.

Para la presente investigación se usó el diagrama de componentes y contenedores para visualizar la organización de los servicios del sistema, sus dependencias y relaciones, como se observa en la figura 1.

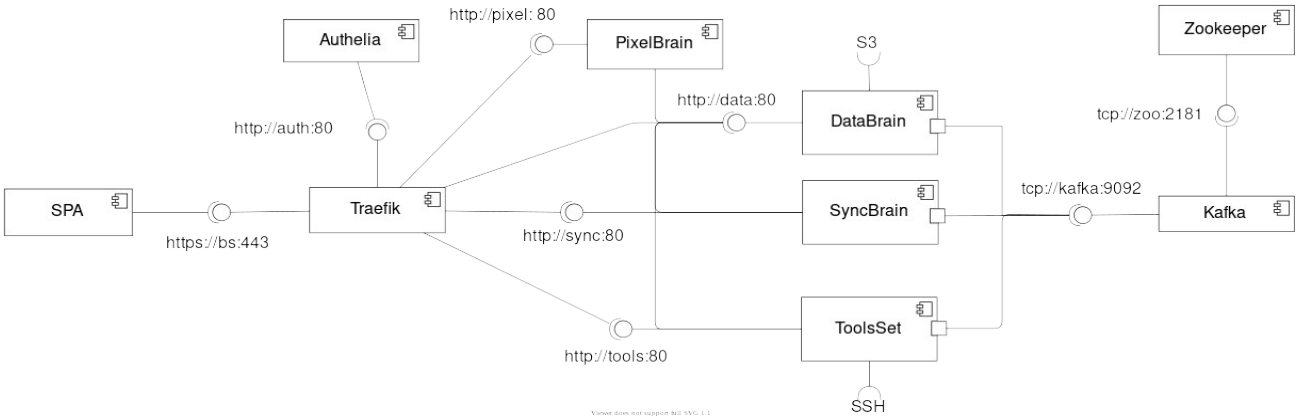

Fig. 1. Vista de servicios usando un diagrama de componentes. Fuente: los autores.

La arquitectura se encuentra distribuida entre 9 servicios lógicos conectados a partir de protocolos de comunicación ligeros como HTTP, SSH, TCP y S3 descritos a continuación:

- Servicios
  - **SPA**: Acrónimo de "Aplicación de una sola página" o *"Single Page Aplication"* por sus siglas en inglés. Es la interfaz gráfica del sistema
  - **Traefik**: Es el enrutador y puerta de enlace encargado de redirigir el tráfico de red a los componentes del sistema. También tiene la capacidad de descubrir servicios en tiempo real.
  - **Authelia**: Es un sistema de autenticación única capaz de integrarse con proxys y puertas de enlace, propicia las funcionalidades de autenticación y autorización. Cuenta con soporte para varios sistemas de proveedores de identidad entre ellos LDAP.
  - **DataBrain**: servicio de almacenamiento de neurodatos, para ello usa una conección a un sistema de objetos compatible con el protocolo S3 e implementa un patrón CQRS ("*Command and Query Responsability Segregation*" por sus siglas en inglés) en su interior.
  - **SyncBrain**: servicio de sincronización encargado de sincronizar los datos de sistemas externos como LORIS con los datos del repositorio.
  - **PixelBrain**: servicio de visualización cuyo objetivo es representar visualmente la información de los datos del repositorio
  - **ToolsSet**: este servicio se encarga de la administración de procesos de ejecución en clústeres y computadoras de alto rendimiento usando los datos del repositorio.
  - **Kafka**: es una plataforma distribuida de flujos de eventos usada con frecuencia para *pipelines* de procesamiento de datos, integración de datos y análisis de flujos. Es el componente central que permite la comunicación de los demás componentes entre sí e incrementa la adaptabilidad del sistema.
  - **ZooKeeper**: es un servicio centralizado para mantener la información de configuración, nombrar, brindar sincronización distribuida y brindar servicios grupales. Todos estos tipos de servicios son utilizados de una forma u otra por aplicaciones distribuidas.
- Protocolos de comunicación:

- **HTTP**: el "Protocolo de Transferencia de Hipertexto" o *"Hyper Text Transfer Protocol"* por sus siglas en inglés es un protocolo de capa de aplicación en el modelo de conjunto de protocolos de Internet para sistemas de información hipermedia distribuidos y colaborativos.
- **TCP**: el "Protocolo de Control de Transmisión" o "*Transmision Control Protocol*" por sus siglas en inglés es uno de los principales protocolos del conjunto de protocolos de Internet. Se originó en la implementación inicial de la red en la que complementó el Protocolo de Internet.
- **SSH**: el "Protocolo de *Shell* Seguro" o "*Secure Shell Protocol*" es un protocolo de red criptográfico para operar servicios de red de forma segura a partir de una red no segura.
- **S3**: la mayoría de los proveedores de almacenamiento de objetos admiten la API de S3 para recuperar y enviar datos al sistema de almacenamiento de objetos. Es una tendencia que las aplicaciones de copia de seguridad, archivado y otras ramas.

Los servicios mencionados son desplegados como unidades independientes de código (ver figura 2) configurados por una "maquinaria" de orquestación que permite despliegues parciales, alta mantenibilidad y escalabilidad.

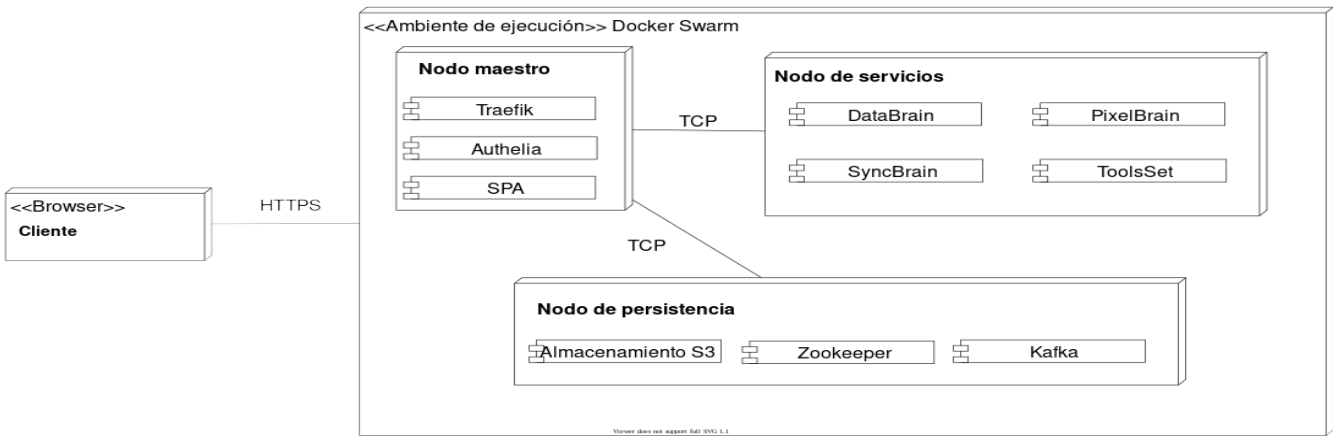

Figura 2. Vista de despliegue de la arquitectura propuesta. Fuente: los autores.

- ***Docker Swarm***: software y servicio de orquestación basado en *docker* y contenedores.
- **Nodo maestro**: entorno de ejecución físico o virtual. Es el nodo encargado de todo la configuración y despliegue del *Docker Swarm*.
- **Nodo de servicios**: entorno de ejecución físico o virtual donde se despliegan los contenedores de servicios.
- **Nodo de persistencia**: entorno de ejecución físico o virtual donde se despliegan los contenedores de persistencias.
- **Cliente**: un navegador web capaz de cargar la aplicación desde el servicio SPA en el nodo maestro.

IV. CONCLUSIONES

Los nuevos estilos arquitectónicos producto de la modernización tecnológica propician el diseño de sistemas más adecuados a las necesidades de los centros de investigación y a los procedimientos ejecutados durante la gestión, procesamiento y análisis de datos de neurociencia.

La arquitectura de microservicios dirigidos por eventos permite diseños adaptables que pueden interoperar entre sí y con sistemas externos ajenos a su propio ambiente de ejecución, útiles para el procesamiento de datos en general y en el caso de la presente investigación para el procesamiento de datos de neurociencia.

La naturaleza distribuida de la arquitectura permite asegurar elementos de calidad como escalabilidad, mantenimiento, reusabilidad e interoperabilidad reduciendo la velocidad en que se acumula deuda técnica en los sistemas.

AGRADECIMIENTOS

La investigación que da origen a los resultados presentados en la presente publicación recibió fondos de la Oficina de Gestión de Fondos y Proyectos Internacionales bajo el código PN305LH013-038.

REFERENCIAS

1.Neuroscience [Internet]. [citado 20 de enero de 2022]. Disponible en: http://www.merriam-webster.-com/medlineplus/neuroscience
2. Principles of Neural Science. McGraw-Hill Education;
3. Ayd FJ. Lexicon of Psychiatry, Neurology and Neurosciences. Lippincott, Williams y Wilkins; 2000. 688 p.
4.Shulman RG. Neuroscience: A Multidisciplinary, Multilevel Field. Oxford University Press; 2013. 59 p.
5. Ogawa H, Okka K. Methods in Neuroethological Research. Springer. 2013;
6. Pasnksepp J. A role for affective neuroscienceïn understanding stress: the case of separation distress circuitry [Internet]. Kluwer Academic; 1990. Disponible en: https://link.springer.com/chapter/10.1007/978-94-009-1990-7_4
7. Cellular neuroscience - Latest research and news [Internet]. [citado 2 de febrero de 2022]. Disponible en: https://www.nature.com/subjects/cellular-neuroscience
8. Computational neuroscience - Latest research and news [Internet]. [citado 2 de marzo de 2022]. Disponible en: https://www.nature.com/subjects/computational-neuroscience
9. Revest P, Longsraff A. Molecular Neuroscience,. Garland Science; 1998.
10. Definition of NEUROCHEMISTRY [Internet]. [citado 2 de septiembre de 2022]. Disponible en: https://www.merriam-webster.com/dictionary/neurochemistry
11. Neuroanatomy - an overview [Internet]. Disponible en: https://www.sciencedirect.com/topics/psychology/neuroanatomy
12. Pampolini N, Giugliano M, Scaini D, Ballerini L, Rauti R. Advances in Nano Neuroscience: From Nanomaterials to Nanotools. Frontiers in Neuroscience [Internet]. 2019;12. Disponible en: https://www.frontiersin.org/article/10.3389/fnins.2018.00953

13. Zhang J, Chen K, Wang D, Gao F, Zheng Y, Yang M. Advances of Neuroimaging and Data Analysis. Frontiers in Neurology [Internet]. 2020;11. Disponible en: https://www.frontiersin.org/article/10.3389/fneur.2020.00257
14. Frontiers in Neuroinformatics [Internet]. Disponible en: https://www.frontiersin.org/journals/neuroinformatics
15. Impact of picture archiving and communication system (PACS) on radiology staff | Elsevier Enhanced Reader [Internet]. [citado 10 de julio de 2022]. Disponible en: https://reader.elsevier.com/reader/sd/pii/S2352914817301958?token=DA40E950B4889256540742EB398F246BFCD41B3EBF132F38B55F277D5EB9EB3401C96882E25788DBE5C680B9BEE162C7&originRegion=eu-west-1&originCreation=20220710205228
16. Insua DR, Oteiza DGU. BigData: Conceptos, tecnologías y aplicaciones. :96.
17. Das S, Zijdenbos AP, Harlap J, Vins D, Evans AC. LORIS: a web-based data management system formulti-center studies. Frontiers in Neuroinformatics. 2012
18. Tarek S, Pierre R, Marc-Etienne R, Nicolas K, Natacha B, Reza A, et al. CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. Frontiers in Neuroinformatics [Internet]. 2014;8. Disponible en: https://www.frontiersin.org/article/10.3389/fninf.2014.00054
19. Falk B, Vogelstein JT. NeuroData's Open Data Cloud Ecosystem.
20. . .Pahl C, Jamshidi P. Microservices: A Systematic Mapping Study. En: Proceedings of the 6th International Conference on Cloud Computing and Services Science [Internet]. Rome, Italy: SCITEPRESS - Science and Technology Publications; 2016 [citado 10 de junio de 2022]. p. 137-46. Disponible en: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005785501370146