Xelhua

sistema agnóstico en la nube para la construcción de soluciones de big data basada en el diseño de servicios de ciencia de datos de alta disponibilidad y tolerante a fallos.

Convocatoria FORDECYT 2019-06 CONACyT

Resumen ejecutivo

Proyecto **41756**

Responsable Técnico:

Dr. José Luis González Compeán

joseluis.gonzalez@cinvestav.mx

Profesor-Investigador

1. Resumen Ejecutivo

El presente reporte describe el entregable 2.1 del Proyecto Número 41756 llamado Xelhua: sistema agnóstico en la nube para la construcción de soluciones de big data basada en el diseño de servicios de ciencia de datos de alta disponibilidad y tolerante a fallos. Este sistema de big data agnóstica está enfocado en la construcción, asistida por el diseño, de servicios de ciencia de datos de alta disponibilidad para la toma de decisiones basada en datos. El sistema Xelhua consta de cuatro componentes principales:

1. Un marco de diseño de alto nivel para la selección de herramientas analíticas y de aprendizaje automático, a través de una malla de servicios acoplados basada en pipelines de procesamiento.
2. Un nuevo modelo de procesamiento basado en el modelo de Extracción-Transformación-Carga (ETL, por sus siglas en inglés) recursivo para convertir automáticamente los diseños de pipelines en estructuras de software agnósticas a la infraestructura.
3. Un modelo novedoso de orquestación para gestionar, de forma transparente, la entrega de datos a lo largo de cada etapa ETL de los pipelines de procesamiento utilizados en los sistemas de ciencia de datos.
4. Un modelo descentralizado de datos para enmascarar de forma transparente la indisponibilidad de algún servicio debido a, por ejemplo, las interrupciones en la nube y la indisponibilidad de las aplicaciones o los datos.

En este reporte también se presentan los contenidos desarrollados con propósito de difusión, así como la documentación técnico-científica de los productos conseguidos por Xelhua en el contexto del proyecto ProNacEs Número 41756.

1.1. Productos académicos:

A continuación, se listan los productos académicos resultantes durante la primera etapa del proyecto:

- Artículos de revista
  - Xel: A cloud-agnostic data platform for the design-driven building of high-availability data science services (2023). Juan Armando Barrón Lugo, Jose luis Gonzalez, Ivan Lopez-Arevalo, Jesus Carretero and Jose L. Martinez-Rodriguez. Future Generation Computer Systems.
- Capítulos de libro:
  - Xelhua: una plataforma para la creación de sistemas de ciencia de datos bajo demanda (https://repositorio-salud.conacyt.mx/jspui/handle/1000/273)

2. Comprometidos CAR

- Módulos para realizar procesos de estadística básica, agrupamiento estadístico, consultas espaciotemporales con operadores condicionales y ejecución de algoritmos de cómputo evolutivo.
- Un esquema de encapsulación y adecuación de algoritmos de análisis, minería y cómputo evolutivo a las estructuras de cripto-aplicaciones.
- Un reporte técnico describiendo los resultados obtenidos.

3. Sistema agnóstico en la nube para la construcción de soluciones de big data basada en el diseño de servicios de ciencia de datos de alta disponibilidad y tolerante a fallos.

Los sistemas de expediente clínico electrónico (SECE) han sido herramientas clave para mejorar los procesos de atención de pacientes. Sin embargo, aún existen áreas de mejora divididas en dos vertientes: i) el cumplimiento de requisitos de interoperabilidad, eficiencia, seguridad, resistencia a fallas y trazabilidad de transacciones, los cuales no son provistos, en su conjunto, por los SECES; ii) la posibilidad de que tanto los SECE, sus fuentes de datos (personal médico, sensores y dispositivos médicos) y la información producida por los SECE (datos históricos) pueda convertirse en una fuente de datos para algoritmos de analítica con soporte de procesos de toma de decisiones.

Ambas áreas de oportunidad representan un desafío, ya que los datos de salud son de alta sensibilidad y su manejo deben sujetarse a las normas oficiales de protección de datos personales, las cuales no son cubiertas, en su totalidad, por los SECE disponibles en México.

Muyal-Ilac es un proyecto de carácter multidisciplinario donde colaborarán investigadores y especialistas en telecomunicaciones, ciencias de datos, informática médica y tecnologías de la información para el diseño y desarrollo de esta plataforma para la gestión, aseguramiento, intercambio y preservación de grandes volúmenes de datos en salud (big data) que proveerá servicios para facilitar a los SECE el cumplimiento de normas oficiales y estándares internacionales de operatividad y seguridad. Esta plataforma contempla tres grupos de servicios:

1. Integración, gestión y compartición de información de pacientes, generada a partir de la práctica clínica, así como la interoperabilidad en el intercambio de datos entre los SECE existentes, sin modificarlos.
2. Esquemas de criptografía de siguiente generación, sistemas de almacenamiento digital, sistemas distribuidos de transmisión/descarga de imágenes radiológicas (PACS) y entrega de contenidos basados en flujos de trabajo de publicación/suscripción creados con tecnologías de perímetro, la nube e internet de las cosas, y
3. Servicios de análisis/estadística de datos que podrán acceder automáticamente a los datos publicados por instituciones a través de investigadores autorizados.

Particularmente, el sistema Xelhua forma parte de la solución al tercer grupo de servicios de la plataforma Muyal-Ilac con el fin de realizar estudios espaciales temporales con mapas de riesgo basados en bases de datos de egresos de algunas enfermedades crónicas.

El diseño e implementación del sistema Xelhua está enfocado en cuatro componentes principales:

1. Un marco de diseño de alto nivel. Permite seleccionar diferentes herramientas de análisis de datos y de aprendizaje automático a partir de una malla de servicios acoplados en pipelines de procesamiento. El sistema Xelhua implementa un servicio de diseño impulsado por datos (figura 1, desarrollo), permitiendo crear pipelines de big data de alto nivel, lo que produce grafos acíclicos dirigidos (DAG , por sus siglas en inglés).
2. Un nuevo modelo de procesamiento de Extracción-Transformación-Carga (ETL , por sus siglas en inglés) recursivo. Permite al sistema Xelhua convertir automáticamente los diseños de pipelines en estructuras de software independientes a la infraestructura, basándose en el DAG producido en la fase de diseño. ETL, es un proceso de integración de datos que extrae, transforma y carga datos de múltiples fuentes a un almacén de datos o a otro repositorio de datos unificado [34]. Este modelo se encarga de encapsular las aplicaciones analíticas de datos, en imágenes de software genéricas agnósticas a la infraestructura, denominadas ABox, que incluyen las dependencias, bibliotecas y sistemas operativos requeridos por las aplicaciones analíticas para ser ejecutadas en una plataforma de contenedores virtuales (figura 1, desarrollo). Estas imágenes también incluyen interfaces de entrada/salida para interconectar diferentes estructuras ABox, creando los pipelines.
3. Un modelo de orquestación para gestionar de forma transparente la entrega y recuperación de datos a lo largo de cada fase de los pipelines de procesamiento. Este modelo garantiza que el intercambio de datos se orqueste siguiendo la dirección de las aristas del DAG (figura 1, ejecución)).
4. Un modelo descentralizado que enmascara automáticamente los incidentes de indisponibilidad de los servicios para reducir los efectos secundarios del bloqueo del proveedor relacionados con los cortes o la indisponibilidad de los datos y la infraestructura. Este modelo se basa en esquemas de gestión de datos y eventos, implementados en un software que se incrusta en las estructuras ABox. Estos esquemas de gestión de eventos cumplen con los requisitos no funcionales (NFR , por sus siglas en inglés) para garantizar el funcionamiento continuo de los servicios de big data mediante la creación de redes P2P (figura 1, operación).

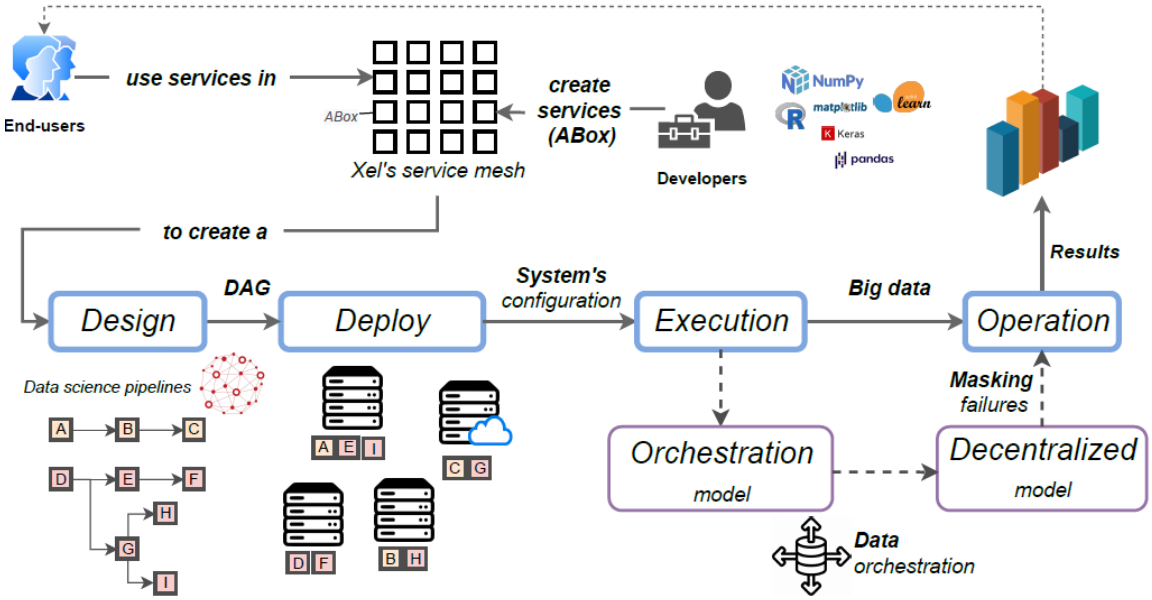

Figura 1. Representación gráfica de la metodología diseñad.

El sistema Xelhua permite construir soluciones de alto nivel a través de un esquema impulsado por el diseño, convirtiendo automáticamente los diseños de pipelines en servicios de ciencia de datos agnósticos y de alta disponibilidad en la nube, desplegables en múltiples infraestructuras para hacer frente a los efectos secundarios del bloqueo del proveedor. En tiempo de ejecución, un motor de orquestación crea flujos de datos continuos, mientras que un modelo descentralizado garantiza las operaciones continuas de estos servicios de big data enmascarando los fallos detectados, como la indisponibilidad de aplicaciones y datos.

Importante: Para más información acerca de Xelhua, refiérase al reporte técnico "Xelhua: sistema agnóstico en la nube para la construcción de soluciones de big data basada en el diseño de servicios de ciencia de datos de alta disponibilidad y tolerante a fallos.".