Revista Cubana de Administración Pública y Empresarial
ISSN 2664-0856 RNPS 2458 / Vol. 6 Núm. 3 / Septiembre-Diciembre (2022) / e249
Disponible en: https://apye.esceg.cu/index.php/apye/article/view/249

Artículo original

ARQUITECTURA DE REFERENCIA DE ECOSISTEMAS DE DATOS BASADA EN DATA MESH & DATA FABRIC

DATA ECOSYSTEMS REFERENCE ARCHITECTURE BASED ON DATA MESH & DATA FABRIC

Tatiana Delgado Fernández * https://orcid.org/0000-0002-4323-9674

Universidad Tecnológica de La Habana "José Antonio Echeverría", La Habana, Cuba

*Autor para dirigir correspondencia: tatiana.delgado@uic.cu

Clasificación JEL: O32, O33, O39

DOI: https://doi.org/10.5281/zenodo.7294747

Recibido: 18/08/2022
Aceptado: 25/10/2022

Resumen

La transformación digital exige cambios acelerados y profundos para aprovechar las tecnologías y los datos en función de hacer más eficaz la toma de decisiones con agilidad y autosostenibilidad. La complejidad de los datos en la era moderna y los silos que se generan a gran escala impulsan la emergencia de nuevos modelos y arquitecturas de gestión de datos que se enfocan a las características intrínsecas de los ecosistemas digitales, caracterizados por las fuertes interrelaciones de diversos actores a lo largo de la cadena de valor, las plataformas como base para interoperar entre ellos y la coevolución de los productos de datos que emanan de fuentes cada vez más heterogéneas. Este artículo propone el diseño de una arquitectura de referencia de ecosistemas de datos basada en las arquitecturas de datos que mejor están soportando la gestión de datos en este complejo escenario: *Data Mesh* y *Data Fabric*, y con el empleo de grafos de conocimiento para la integración. Como método se emplea un análisis de la literatura más reciente sobre gestión y arquitecturas de datos para extraer los principios y componentes arquitectónicos que se emplean en el diseño de tal arquitectura de referencia. Se obtiene una representación abstracta de arquitectura de referencia de ecosistemas de datos, cuyo modelo operacional se verifica teóricamente. La misma es el punto de partida de futuras investigaciones que se encaminarán

Publicado por Escuela Superior de Cuadros del Estado y del Gobierno, La Habana, Cuba

Escuela Superior de Cuadros del Estado y del Gobierno

hacia su implementación en casos de uso reales y el modelado organizacional relativo a los roles de los actores que se involucran en el ecosistema reflejado en la propia arquitectura.

Palabras clave: arquitectura de datos, ecosistema, Data Mesh, Data Fabric, grafos de conocimientos

Abstract

Digital transformation requires rapid and profound changes to take advantage of technologies and data in order to make decision-making more effective with agility and self-sustainability. The complexity of data in the modern era and the silos that are generated at big scale drive the emergence of new data management models and architectures that focus on the intrinsic characteristics of digital ecosystems, characterized by the strong interrelationships of various actors through along the value chain, the platforms as the basis for interoperating with each other and the co-evolution of data products that emanate from increasingly heterogeneous sources. This article proposes the design of a reference architecture for data ecosystems based on the data architectures that are best supporting data management in this complex scenario: Data Mesh and Data Fabric, and with the use of knowledge graphs for the integration. As a method, an analysis of the most recent literature on data management and architectures is used to extract the principles and architectural components that are used in the design of such a reference architecture. An abstract representation of the reference architecture of data ecosystems is obtained, whose operational model is theoretically verified. It is the starting point for future research that will be directed towards its implementation in real use cases and organizational modeling related to the roles of the actors involved in the ecosystem reflected in the architecture itself.

Keywords: data architecture, ecosystem, Data Mesh, Data Fabric, knowledge graphs

Introducción

Los datos tienen un papel y un valor cada vez más importantes para facilitar la toma de decisiones. El volumen, la variedad, la velocidad, la veracidad, como requisito de calidad, y el valor que suponen los datos modernos suelen emplearse para definir el concepto *Big data*, un término que más que datos grandes o masivos, caracteriza su complejidad y el cambio paradigmático que ha venido ocurriendo en las arquitecturas que los gestionan.

Los llamados datos analíticos se están convirtiendo en un componente cada vez más crítico del panorama tecnológico. Son la base para visualizaciones e informes que brindan información sobre un negocio u organización. Además, se utilizan para entrenar modelos de aprendizaje automático que aumentan el negocio con inteligencia basada en datos. Es el ingrediente esencial para que las organizaciones pasen de la intuición y la toma de decisiones guiada por el instinto a la adopción de medidas basadas en observaciones, y predicciones soportadas en datos. Permite un cambio tecnológico de algoritmos basados en reglas, diseñados por humanos, a modelos de aprendizaje automático.[1]

En este nuevo escenario, se hace más palpable el desafío de los "silos de datos" por la naturaleza cada vez más heterogénea de los mismos. Un silo de datos significa que los datos no son tan accesibles como

deberían ser o tal vez no para los equipos que no son los que los generan. Si se requiere una gran cantidad de tiempo solo para decodificar los datos para que sean traducibles a otro equipo, es probable que haya uno o más silos de datos en la organización. Los silos de datos surgen de problemas estructurales (muchas capas de separación entre equipos), culturales (es decir, mantener los datos separados, en lugar de trabajar juntos) y tecnológicos (es probable que las aplicaciones no estén diseñadas para integrarse juntas).[2]

Para enfrentar estos desafíos, los modelos y arquitecturas que soportan la gestión de los datos están cambiando. Una de las arquitecturas más mencionadas en los círculos de avanzada en este entorno es *Data Mesh* o tejido de datos, considerado un enfoque sociotécnico descentralizado para compartir, acceder y administrar datos analíticos en entornos complejos y de gran escala, dentro o entre organizaciones. Se basa en cuatro principios fundamentales: propiedad del dominio, datos como producto, plataforma de autoservicio de datos y gobernanza computacional federada.[1]

Otra arquitectura emergente que está posicionándose en el escenario de gestión de datos es *Data Fabric* (DF) que, en general, se puede definir como un conjunto de principios de gestión de datos, prácticas rectoras, comunidades y estándares que pueden "... optimizar el acceso a los datos distribuidos de una organización y curarlos y organizarlos de manera inteligente para la entrega de autoservicio."[3] Es un sistema que proporciona una arquitectura unificada para administrar y proporcionar datos. Generalmente, se realizan como sistemas distribuidos orientados a servicios donde los conjuntos de servicios proporcionan interfaces consistentes y mecanismos para acceder a datos y capacidades de almacenamiento.[4]

Considerando las diversas y múltiples interrelaciones que surgen entre los conjuntos de datos de diferentes dominios y los también diversos actores que los gestionan, frecuentemente se describen estos escenarios como ecosistemas digitales,[5] dentro de los cuales, los ecosistemas de datos emergen con especial fuerza, dada la importancia crucial que alcanza la gestión integrada de datos para tomar decisiones con mayor eficacia y basada en contexto.

Un ecosistema de datos se puede definir como: un conjunto de redes, compuestas por actores autónomos que directa o indirectamente consumen, producen o proporcionan datos y otros recursos relacionados (por ejemplo, software, servicios e infraestructura). Cada actor desempeña uno o más roles y está conectado a otros actores a través de relaciones, de tal manera que la colaboración y competencia de los actores promueve la autorregulación del ecosistema de datos.[6]

El objetivo de este artículo es proponer una arquitectura de referencia de ecosistemas de datos, basada en la fusión de componentes de las emergentes arquitecturas *Data Mesh* y *Data Fabric*, para ofrecer un modelo abstracto del cual se puedan instanciar diferentes arquitecturas de ecosistemas de datos de distintos dominios, empresas, e incluso, para eliminar los silos de datos interinstitucionales, y alcanzar un nivel de integración de datos del gobierno a nivel central, como parte de la implementación de las políticas de transformación digital.

Materiales y Métodos

Para definir la arquitectura de referencia de ecosistemas de datos, se emplea una metodología de investigación híbrida donde se integran métodos de análisis de literatura para sustentar científicamente la propuesta, junto a otros métodos de modelado de arquitecturas de datos modernas, que den respuesta a una gestión eficaz de los datos complejos que se generan en la era de la transformación digital. La **Figura 1** muestra el esquema de las fases que guiaron metodológicamente el desarrollo de la investigación.

Figura 1. Metodología para definir la arquitectura de referencia de ecosistemas de datos
Fuente: elaboración propia.

La primera fase corresponde con el análisis de la literatura, para lo que fue usado Google académico, debido a su versatilidad, al cubrir una amplia variedad de publicaciones, como artículos, libros, actas de congreso, tesis, y otros materiales. Se incluyen, además, algunas fuentes de la llamada literatura gris, en este caso, fuentes electrónicas provenientes de líderes globales en el tema de gestión de datos, que están marcando pautas en las arquitecturas de datos, en particular en relación con las emergentes arquitecturas *Data Mesh* y *Data Fabric*. El objetivo de aplicar este método es revelar los principios y características distintivas de estas arquitecturas que por ser tan disruptivas se encuentran pocas evidencias científicas de su implementación. Sin embargo, se identifican como tendencias en los informes de consultoras mundiales, como Gartner, y proveedores como IBM y Microsoft. Una vez que se analizan las arquitecturas, cuyos componentes serán evaluados para reutilizar en la de referencia, se pasa a establecer el marco conceptual de ecosistemas de datos, que ocupa otro cuerpo de conocimientos, aunque muy interrelacionado con el primero.

La segunda fase es como tal el diseño de la arquitectura de referencia de ecosistemas de datos. Con los componentes arquitectónicos de *Data Mesh* y *Data Fabric*, que resultaron del estudio de las arquitecturas seleccionadas en la primera fase, se diseña la nueva arquitectura de referencia, manteniendo especial cuidado en los principios heredados de sus antecesoras y de los propios ecosistemas de datos.

Resultados

Data mesh o tejido de datos

Los cuatro principios que sustenta la arquitectura lógica y el modelo operativo de un tejido de datos son: (1) propiedad de datos descentralizada orientada al dominio, (2) datos como producto, (3) plataforma de datos de autoservicio, y (4) gobierno computacional federado.[1] Estos principios se describen a continuación:

1. Principio de propiedad del dominio. Descentraliza la propiedad de los datos analíticos a los dominios de negocio más cercanos a los datos, ya sea la fuente de los datos o sus principales consumidores. Descompone los datos (analíticos) de forma lógica y en función del dominio de negocio que representan, y gestiona el ciclo de vida de los datos orientados al dominio de forma independiente. Alinea arquitectónica y organizativamente datos de negocio, tecnológicos y analíticos. Existen tres arquetipos de datos orientados al dominio: datos de dominio alineados con la fuente, datos de dominio agregados y datos analíticos.
2. Principio de datos como producto. Con este principio en vigor, los datos orientados al dominio se comparten como un producto directamente con los usuarios de datos: analistas de datos, científicos de datos, etc. Cada producto de datos es autónomo y su ciclo de vida y modelo se gestionan independientemente de los demás. Los datos como producto introducen una nueva unidad de arquitectura lógica llamada *quantum* de datos, que controla y encapsula todos los componentes estructurales necesarios para compartir datos como un producto (datos, metadatos, código, política y declaración de dependencias de infraestructura) de forma autónoma. Obtiene un mayor valor de los datos al compartir y usar datos más allá de los límites de la organización.
3. Principio de la plataforma de autoservicio de datos. Este principio conduce a una nueva generación de servicios de plataforma de datos de autoservicio que permiten a los equipos multifuncionales de dominios compartir datos. Los servicios de la plataforma se centran en eliminar la fricción del viaje de extremo a extremo del intercambio de datos, desde la fuente hasta el consumidor. Los servicios de la plataforma gestionan el ciclo de vida completo de los productos de datos individuales. Gestionan un tejido fiable de productos de datos interconectados. Proporcionan experiencias a nivel de tejido, como mostrar el grafo de conocimiento emergente y el linaje a través del tejido. La plataforma agiliza la experiencia de los usuarios de datos para descubrir, acceder y utilizar productos de datos. Asimismo, agiliza la experiencia de los proveedores de datos para crear, implementar y mantener productos de datos.
4. Principio de gobernanza computacional federada. Este principio crea un modelo operativo de gobierno de datos basado en una estructura federada de toma de decisiones y de responsabilidad, con un equipo compuesto por representantes de dominio, plataforma de datos y expertos en la materia: legal, conformidad, seguridad, etc. El modelo operativo crea un incentivo y estructura de rendición de cuentas que equilibra la autonomía y la agilidad de los dominios, con la interoperabilidad global del tejido. El modelo de ejecución de gobierno se basa en gran medida en la codificación y automatización de las políticas en un nivel detallado, para cada producto de datos, a través de los servicios de la plataforma.

El tejido de datos proporciona la integración flexible y resiliente de las fuentes de datos entre distintas plataformas y usuarios comerciales, para que estén disponibles desde cualquier lugar donde se necesiten e independientemente de dónde se alojen.[7]

La plataforma multiplano[1] del tejido de datos permite distinguir entre diferentes clases de servicios de plataforma en función de su ámbito de operación sin imponer una estratificación estricta. Los tres planos de la plataforma incluyen:

- Plano de infraestructura de datos. Servicios atómicos para aprovisionar y administrar recursos físicos como almacenamiento, orquestación de canalizaciones, cómputo, etc.
- Plano de experiencia del producto de datos. Servicios de abstracción de nivel superior que operan directamente con un producto de datos y permiten a los productores y consumidores de productos de datos crear, acceder y proteger un producto de datos, entre otras operaciones que se ejecutan en un producto de datos.
- Plano de experiencia del tejido. Servicios que operan en un tejido de productos de datos interconectados, como la búsqueda de productos de datos y la observación del linaje de datos entre ellos.

Los consumidores de la plataforma (desarrolladores de productos de datos, consumidores, propietarios, la función de gobierno, etc) pueden acceder directamente a todos estos planos.

La **Figura 2** ofrece una vista de los tejidos de datos como una capa integrada de datos conectados.

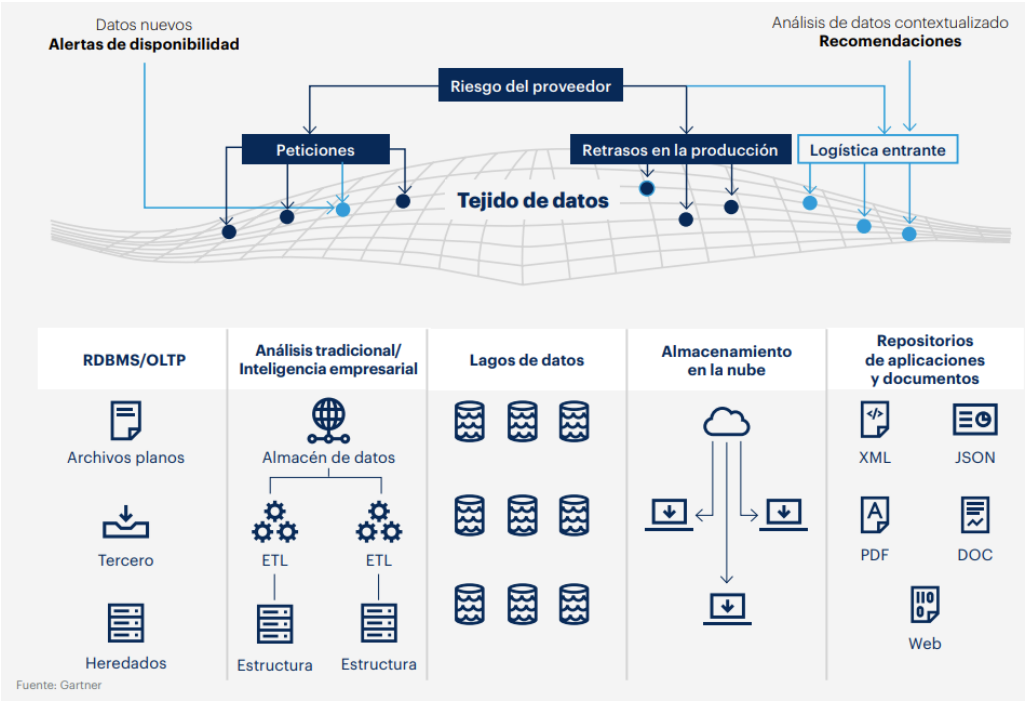

Figura 2. Tejido de datos como capa integrada de datos conectados

Fuente: Gartner[7]

Este enfoque de Gartner sobre tejido de datos es más cercano al de la arquitectura *Data Fabric*, porque identifica de forma más consciente esta capa tecnológica integrada, que generalmente es resuelta en forma de grafos de conocimientos.

Data Fabric

Data Fabric que en español también se puede traducir como tejido de datos y para no confundir con el anterior enfoque se llamará por su término o siglas en inglés (DF), constituye una arquitectura de información y una plataforma para la gestión de datos y la integración a nivel de datos, y proporciona interfaces, API y servicios para la integración y comunicación de los sistemas involucrados. Desde una perspectiva de nivel de sistemas, DF puede verse como un sustrato de comunicación que proporciona un mecanismo unificado para el acceso y la manipulación de datos para todas las herramientas del proyecto.[4]. Esta arquitectura de datos es más bien un marco que permite la implementación automática e inteligente de extremo a extremo de múltiples canales de datos, así como entornos de nube.[3]

La naturaleza distribuida de las DF permite la escalabilidad, la implementación flexible y la adaptación del sistema y, a menudo, se aprovecha para, por ejemplo, facilitar la integración de sistemas a través de los límites organizacionales o combinar el uso de recursos locales y basados en la nube. Si bien se diseñaron principalmente como sustratos para la administración de datos y la comunicación de sistema a sistema, los DF también pueden exponer interfaces y herramientas a los usuarios finales para facilitar el desarrollo de mecanismos para administrar, buscar y analizar datos de manera conveniente.[4]

Dado que los datos distribuidos de una organización evolucionarán con el tiempo, tanto en contenido como en escala y formato, es importante contar con un enfoque flexible y escalable. Estos conceptos de evolución y escalabilidad son fundamentales en las arquitecturas *Data Fabric*. Un modelo DF siempre debe esforzarse por cooperar con las comunidades y los grupos de trabajo debido al valor inherente de los datos y servicios conectados. Parte de esa cooperación puede resultar y resulta en estándares y enfoques publicados formalmente. También hay que considerar principios de diseño, como:

- Disponibilidad de datos, referido a que los datos deben ser intuitivos.
- Valor de los datos, es decir, tienen un valor intrínseco.
- Datos conectados, que significa que son inherentemente más valiosos cuando están conectados.
- Los datos deben ser FAIR, conocido por sus siglas en inglés que en español significan encontrables, accesible, interoperable y reutilizables.
- Armonía entre datos, servicios y software. Es recomendable usar ontologías y varios otros enfoques semánticamente expresivos, tanto técnica como conceptualmente.
- Aprender de sus datos, es decir si se capturan y conectan correctamente, puede aprender mucho sobre sus datos, lo cual proporcionarán valor a su esfuerzo y misión, en general. Se recomienda usar inteligencia artificial para tratar de aprender patrones que podrían ser útiles a un nivel más amplio, usar análisis estadísticos para determinar las formas más eficientes de resolver problemas y utilizar flujos de trabajo predictivos basados en el conocimiento actual, entre otras técnicas.

Con el fin de aumentar la salud de los datos, *Data Fabric* ofrece capacidades integradas de calidad de los datos, preprocesamiento de datos y gobernanza de la información que están habilitadas por el aprendizaje automático y la automatización mejorada.[8]

DF se resume como un diseño de gestión de datos que permite la integración y el intercambio de datos entre fuentes de heterogéneas, para lograr integración flexible, reutilizable y aumentada de datos, que utilizan grafos de conocimiento, semántica y aprendizaje automático/inteligencia artificial, en metadatos activos, para respaldar un acceso y uso compartido de datos de forma más rápida y, en algunos casos, automatizado, independientemente de las opciones de implementación, casos de uso (operativos o analíticos) o enfoques arquitectónicos.[7]

La integración a nivel de datos en *Data Fabric* ocurre con frecuencia a través de los grafos de conocimiento (KG). Un KG es un modelo conceptual de un dominio de conocimiento, en este caso el diseño de su producto y su proceso de creación. Los expertos en dominios usan un KG de este tipo para describir y resolver problemas relacionados con el dominio, utilizando sus conceptos del mundo real, el vocabulario y las relaciones entre estos conceptos. El KG no necesariamente debe contener todos los datos disponibles en una organización o el dominio que represente. Esto sería indeseable y, por lo general, incluso poco realista. En su lugar, existen varias opciones para relacionar KG y datos y, por lo general, un KG reúne estos aspectos, equilibrando flexibilidad, costo y actuación.[4] Entre estas opciones se encuentran:

- Individuos directos. Un KG puede contener a todos sus individuos dentro de su infraestructura. Esta es la forma tradicional de construir un KG, que contiene todos los conceptos e individuos.
- KG Virtual. Si bien un KG puede contener a los individuos por sus conceptos, no es necesario que los contenga físicamente en todos los casos. Un KG virtual obtiene dinámicamente algunos de sus individuos de otros tipos de almacenamiento y se los proporciona al usuario que lo solicita. Esto hace que se escalen mejor cuando aumenta el número de individuos.
- KG de referencia. A veces no se desea almacenar individuos en el KG y simplemente acceder a ellos a través del KG. El mero tamaño de los datos relacionados puede ser prohibitivo, o un KG puede no ser adecuado para representar a esas personas. Esto es, por ejemplo, aplicable a los datos de series temporales, que son muy frecuentes en los escenarios de producción del proyecto. Sin embargo, un KG no tiene la estructura adecuada para manejar de manera eficiente cantidades masivas de mediciones basadas en el tiempo para una gran cantidad de propiedades físicas operativas. En su lugar, el KG debe manejar dichos valores como hojas en su estructura: en lugar de contener los valores como individuos y apuntar a la entrada de acceso a datos adecuada del *Data Fabric*.

Data Fabric y *Data Mesh* proporcionan arquitecturas para acceder a los datos a través de múltiples tecnologías y plataformas. Se diferencian en que la primera está centrada en la tecnología, mientras que el tejido de datos se enfoca más en el cambio organizacional. Se pueden mezclar para aprovechar las ventajas de ambos, porque existe una gran compatibilidad entre ellas. *Data Fabric* le da más sentido al cómo se integran los datos de forma armónica al proporcionar explícitamente la variante de grafos de

conocimientos. Esto será particularmente aprovechado en la arquitectura de referencia que se diseña en esta investigación.

Ecosistemas de datos

La metáfora de los ecosistemas se ha utilizado para describir múltiples y variadas interrelaciones entre muchos actores e infraestructura que contribuyen a la creación de un recurso, por ejemplo, negocio, servicio o software.[9] En este sentido, los ecosistemas presentados van más allá de las cadenas de valor tradicionales y la estructura industrial al tener tres características principales: red, plataforma y coevolución. La primera característica establece la existencia de una red flexible de actores, incluidos desarrolladores, proveedores, revendedores y proveedores de tecnología e infraestructura. Todos los actores están comprometidos con la producción de valor o la extracción de valor del ecosistema. La segunda característica es una "plataforma" (por ejemplo, servicios, herramientas o tecnologías) que los actores del ecosistema pueden utilizar para generar beneficios. La plataforma permite que diferentes actores contribuyan al ecosistema y da como resultado un conjunto de productos o servicios. Finalmente, el ecosistema permite que los actores y productos de datos evolucionen conjuntamente, es decir, ser parte de un ecosistema que exige colaboración y conexión entre diferentes actores en diferentes campos de especialización y conocimiento, y entre los artefactos que generan. Al mismo tiempo, ser parte del ecosistema permite a los actores tener acceso entre sí, como proveedores, innovadores o solucionadores de problemas, ya sea que trabajen de forma independiente o dentro de organizaciones de investigación, organizaciones privadas o públicas.[6]

Por lo tanto, un ecosistema de datos puede verse como otra instancia de un ecosistema digital.[5] Además, un ecosistema de datos puede concebirse como parte de múltiples tipos de ecosistemas organizados en torno a empresas, recursos y productos proporcionados por diferentes actores. Los objetivos más amplios de innovación y creación de valor se traducen en términos más específicos relacionados con cada contexto de ecosistema específico. En particular, los datos se pueden utilizar para respaldar negocios, brindar innovación, promover la transparencia para los gobiernos, validar la investigación y muchos otros objetivos. Además de estar interconectados, los límites entre un ecosistema de datos y otros ecosistemas son difíciles de definir. Por ejemplo, un ecosistema de datos puede implicar ecosistemas de software sobre la red de actores involucrados en el desarrollo y suministro de software relacionado con datos.[9] También en la administración pública surgen ejemplos de ecosistemas de datos para el Gobierno.[10]

Los marcos de trabajo y las arquitecturas de los ecosistemas de datos están aún en su infancia y aunque hay algunas interesantes propuestas,[11-14] queda mucho espacio para investigar en busca de un modelo de gestión optimizada de datos complejos, en ambientes distribuidos, como sistemas socio-técnicos, y de forma interoperable e integrada.

Los mecanismos de gestión y arquitecturas de datos son componentes taxonómicos de la transformación digital,[15] que facilitan su adopción. En este artículo, se sitúa en el centro a los ecosistemas de datos para proponer una arquitectura que sea capaz de respaldar las implementaciones tecnológicas y organizacionales de este paradigma, cuya esencia misma está basada en ecosistemas digitales.

Arquitectura de ecosistemas de datos basada en Data Mesh y Data Fabric

Para diseñar la arquitectura de referencia de ecosistema de datos, se reutilizan los cuatro principios de *Data Mesh*[1]: propiedad del dominio, datos como producto, plataforma de autoservicio de datos y gobernanza computacional federada. La **Figura 3** presenta la Arquitectura de referencia de ecosistemas de datos propuesta.

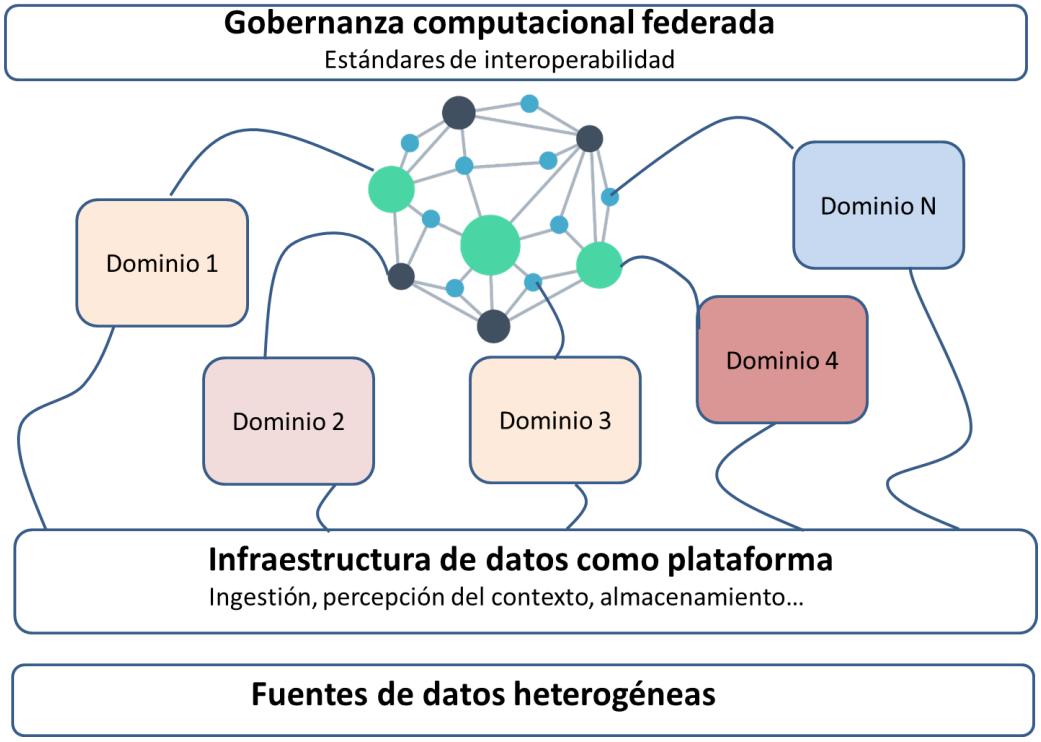

Figura 3. Arquitectura de referencia de ecosistemas de datos

Fuente: elaboración propia

Se asumen los grafos de conocimiento (KG) como el componente arquitectónico idóneo para lograr la integración entre los diferentes productos de datos que se sirven en los dominios.

Por lo general, las instancias de los conceptos de un grafo se consideran los "datos" del KG. Sin embargo, eso no significa que el KG deba contener todos los datos disponibles en una organización. En su lugar, existen varias opciones para relacionar KG y datos, equilibrando flexibilidad, costo y actuación.[4] Entre estas opciones se encuentran:

- KG materializado con individuos directos. Un KG puede contener a todos sus individuos dentro de su infraestructura. Esta es la forma tradicional de construir un KG.[16-18]
- KG Virtual: Si bien un KG puede contener a los individuos por sus conceptos, no es necesario que los contenga físicamente en todos los casos. Un KG virtual obtiene dinámicamente algunos

de sus individuos de otros tipos de almacenamiento y se los proporciona al usuario que lo solicita. Esto hace que se escalen mejor cuando aumenta el número de individuos.[20-21]

- KG de referencia: a veces no se desea almacenar individuos en el KG y simplemente acceder a ellos a través del KG. El mero tamaño de los datos relacionados puede ser prohibitivo, o un KG puede no ser adecuado para representar a esos individuos. Esto es, por ejemplo, aplicable a los datos de series temporales, que son muy frecuentes en los escenarios de producción del proyecto. Sin embargo, un KG no tiene la estructura adecuada para manejar de manera eficiente cantidades masivas de mediciones basadas en el tiempo para una gran cantidad de propiedades físicas operativas. En su lugar, el KG debe apuntar a la entrada de acceso a datos adecuada del *Data Fabric*.

La construcción del grafo está asociada a alguna de estas variantes, pero también a la naturaleza del producto de datos de cada dominio. En la literatura hay metodologías detalladas para la construcción de grafos de conocimientos.[17] Asimismo, se presentan implementaciones en dominios específicos, como, por ejemplo, un estudio llevado a cabo para el diseño de grafos en un escenario relacionado con análisis epidemiológicos de la Covid 19.[16] Otros ejemplos, en el caso de grafos de conocimientos empresariales, están siendo más frecuentemente abordados en el modo directo[18-19] y como KG virtuales.[20-21] Sin embargo, en ninguno de estos casos se trabajan las capas de infraestructuras y de productos de datos por dominio como se muestra en la arquitectura propuesta.

Discusión

La arquitectura de referencia de ecosistemas de datos se verifica a partir de comprobar que en ella se integran las tres características básicas de los ecosistemas, descritas anteriormente: red, plataforma y coevolución, y también con el análisis de cómo refleja los componentes arquitectónicos de las arquitecturas de datos que le dan origen: *Data Mesh* y *Data Fabric*.

Las características presentes en un ecosistema se encuentran explícitamente concebidas, primero, en su funcionamiento como red, con el componente arquitectónico KG aportado por el modelo de arquitectura *Data Fabric*, que permite enlazar los datos y sus propiedades sin considerar las aplicaciones de cada dominio, sino utilizando los datos como producto; mientras que su naturaleza de plataforma se refleja en la incorporación de la capa de infraestructura, que sigue la filosofía de plataforma autoservida de *Data Mesh*.

La capa de infraestructura de datos como plataforma es la encargada de la ingesta de los datos de fuentes heterogéneas, mediante ETL-extraer, transformar y cargar (para datos transaccionales que se gestionarán en dominios de almacenes de datos), o ELT -extraer, cargar y transformar (en el caso de datos complejos a los que se aplican mecanismos de gestión basados en *Big Data*) o cualquier otro mecanismo alineado a las fuentes.

Los dominios son responsables de generar el producto de datos, como se define en la arquitectura de planos de *Data Mesh.*[1]. El plano de experiencia del tejido optimiza la experiencia de las personas que necesitan operar, gobernar y consultar el tejido como un todo. En el caso de la arquitectura de referencia

de ecosistemas de datos presentada, se aprovecha la capacidad específica de los servicios de datos que ofrece cada dominio para servir a los grafos que enlazan los datos a nivel integrado, bajo la filosofía de *Data Fabric*. A la vez, se consideran las políticas y estándares recomendados en la capa de gobernanza computacional federada. Por ejemplo, los miembros del equipo de gobernanza y los propietarios de productos de datos, que trabajan al interior de los dominios, interactúan con los servicios en este plano para evaluar el estado actual de las políticas, monitorear el estado operativo general del tejido y buscar productos de datos existentes.[1] También lo utilizan los consumidores y proveedores de productos de datos en escenarios en los que necesitan trabajar con una colección de productos de datos, como búsquedas y recuperación de datos. Para la arquitectura de referencia propuesta, la consulta a los datos integrados se realiza a través de las herramientas asociadas a grafos de conocimiento, usando SPARQL u otras aplicaciones que embeben su funcionalidad. El plano de experiencia del producto de datos está optimizado para la entrega y el consumo de productos de datos mediante API y a través de grafos de conocimiento cuando se integran varios datos de distintos dominios.

Implicaciones teóricas de la investigación

La arquitectura de ecosistema de datos, al igual que *Data Mesh*, exige un cambio fundamental en la forma en que se administran, usan y se consumen los datos analíticos, tal como se describe:

- El modelo de propiedad de datos descentralizado empuja la propiedad y la responsabilidad de los datos a los dominios de negocio desde donde se producen o se utilizan los datos, primando un modelo federado con políticas computacionales integradas en los nodos del tejido.
- Los datos se sirven como productos, lo cual aprovecha mejor las características intrínsecas de cada fuente de datos, mientras son servidos acorde a la forma en que mejor satisfacen la experiencia del consumidor.
- Arquitectónicamente, se pasa de recopilar datos en almacenes y lagos monolíticos a conectar datos a través de un tejido distribuido de productos de datos a los que se accede a través de protocolos estandarizados, mientras tecnológicamente, las soluciones tratan los datos y el código que los mantiene como una unidad autónoma activa.

Este artículo, cuya principal contribución es el propio diseño de una arquitectura de referencia de ecosistemas de datos que se verifica respecto a los principios de aquellas arquitecturas que le dan origen, también presenta limitaciones. La más importante de ellas es que se basa en un enfoque teórico, por lo que futuras investigaciones deberán abordar métodos empíricos y basados en casos que implementen la arquitectura de referencia para ecosistemas de datos en dominios reales. Por otra parte, hay que seguir investigando en los roles de los actores que involucran un ecosistema de datos y reflejarlo en la arquitectura para que pueda ser implementada de forma gobernable.

Conclusiones

Los ecosistemas de datos se están ratificando como el mejor modelo de representar las múltiples interconexiones entre actores diferentes que da como resultado un conjunto de productos o servicios que se generan a partir de los datos también interconectados. Opuesto a los silos de datos, los ecosistemas de

datos garantizan interoperabilidad e integración a nivel de datos. Deben ser construidos sobre arquitecturas flexibles y escalables, que faciliten la implementación automática e inteligente de extremo a extremo de múltiples canales de datos.

Las tendencias en torno a los marcos para construir tales arquitecturas apuntan a *Data Mesh/Data Fabric*. La arquitectura de referencia de ecosistemas de datos presentada a nivel abstracto en este artículo hereda los principios y características de estos marcos para contemplar un modelo de propiedad de datos descentralizado, donde los dominios de negocio se ocupan de proveer productos de datos al ecosistema, para el consumo de cualquier otro actor/dominio o para ser integrado en el tejido (grafo) desde donde se consumen contextualizados a nivel holístico.

Los trabajos futuros en esta dirección estarán encaminados a realizar pruebas de concepto que validen la arquitectura de referencia y a incorporar el análisis organizacional y las implicaciones prácticas de su adopción.

Referencias bibliográficas

1. Dehghani Z. Data Mesh: Delivering Data-Driven Value at Scale (1.ed - preview version), O'Reilly Media, Inc. 2022. [Consultado 5 septiembre de 2022]. Disponible en: https://www.oreilly.com/library/view/data-mesh/9781492092384/.
2. Fortney J, McDonnell M, Johnson D, Chalk S. Data Fabric and Data as a" First Class Citizen"; 2022. [Consultado 1 spetiembre de 2022]. Disponible en: http://dx.doi.org/10.13140/RG.2.2.14510.18240
3. IBM, "Data fabric," 2021. [Online]. Available: https://www.ibm.com/analytics/data-fabric
4. Östberg PO, Vyhmeister E, Castañé GG, Meyers B, Van Noten J. Domain Models and Data Modeling as Drivers for Data Management: The ASSISTANT Data Fabric Approach. IFAC-PapersOnLine. 2022 Jan 1;55(10):19-24. [Consultado 4 septiembre de 2022]. Disponible en: https://doi.org/10.1016/j.ifacol.2022.09.362
5. Delgado T. Una arquitectura de Ecosistemas de Datos Espaciales. XVI Convención y Feria INFORMATICA 2016: Conectando sociedades 2016; 1-6. ISBN 978-959-289-122-7.
6. de Oliveira EF, Silveira MS. Open government data in Brazil a systematic review of its uses and issues. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age 2018 May 30:1-9. https://doi.org/10.1145/3209281.3209335).
7. Gartner. Understand the role of Data Fabric. Guides for Effective Business Decision Making; 2022. [Consultado 21 agosto de 2022]. Disponible en: https://www.gartner.com/en/publications/essential-guide-to-data-fabric.
8. Liu CM, Badigineni M, Lu SW. Adaptive Blocksize for IoT Payload Data on Fabric Blockchain. In2021 30th Wireless and Optical Communications Conference (WOCC) IEEE. 2021 Oct; 7: 92-96). [Consultado 2 agosto de 2022]. Disponible en: http://doi.org/10.1109/WOCC53213.2021.9602935.
9. Farias VG, Santos R, Wiese I, Serebrenik A, Constantinou E. Investigating Power Relations in Open Source Software Ecosystems. InAnais Estendidos do XII Congresso Brasileiro de Software: Teoria e Prática 2021 Sep 27 (pp. 53-59). SBC. [Consultado 23 julio de 2022]. Disponible en: https://doi.org/10.5753/cbsoft_estendido.2021.17282

10. Shah SI, Peristeras V, Magnisalis I. Government big data ecosystem: definitions, types of data, actors, and roles and the impact in public administrations. ACM Journal of Data and Information Quality. 2021 May 6;13(2):1-25. [Consultado 13 agosto de 2022]. Disponible en: https://doi.org/10.1145/3425709
11. Hernandez-Almazan JA, Chalmeta R, Roque-Hernández RV, Machucho-Cadena R. A Framework to Build a Big Data Ecosystem Oriented to the Collaborative Networked Organization. Applied Sciences. 2022 12;12(22):11494. [Consultado 5 noviembre de 2022]. Disponible en: https://doi.org/10.3390/ app122211494.
12. Herrera F, Sosa R, Delgado T. GeoBI and big VGI for crime analysis and report. In2015 3rd International Conference on Future Internet of Things and Cloud 2015 Aug 24 (pp. 481-488). IEEE. [Consultado 12 julio de 2022]. Disponible en: https://doi.org/10.1109/FiCloud.2015.112
13. Orenga-Roglá S, Chalmeta R. Framework for implementing a big data ecosystem in organizations. Communications of the ACM. 2018 Dec 19;62(1):58-65. [Consultado 21 julio de 2022]. Disponible en: https://doi.org/10.1145/3210752
14. Singh KN, Behera RK, Mantri JK. Big data ecosystem: review on architectural evolution. Emerging Technologies in Data Mining and Information Security. 2019:335-45. [Consultado 1 agosto de 2022]. Disponible en: https://doi.org/10.1007/978-981-13-1498-8_30
15. Fernández TD. Taxonomía de transformación digital. Revista Cubana de transformación digital. 2020;1(1):4-23. [Consultado 15 agosto de 2022]. Disponible en: https://rctd.uic.cu/rctd/article/view/62.
16. Delgado T, Stuart ML, Delgado M. Grafos de conocimiento para gestionar información epidemiológica sobre COVID-19. Revista Cubana de Información en Ciencias de la Salud. 2021 Dec;32(4). [Consultado 12 agosto de 2022]. Disponible en: http://rcics.sld.cu/index.php/acimed/article/view/1686.
17. Hogan A, Blomqvist E, Cochez M, d'Amato C, de Melo G, Gutierrez C, Gayo JE, Kirrane S, Neumaier S, Polleres A, Navigli R. Knowledge graphs. arXiv preprint arXiv:2003.02320. 2020; Mar 4. [Consultado 2 agosto de 2022]. Disponible en: https://arxiv.org/abs/2003.02320.
18. Gomez-Perez JM, Pan JZ, Vetere G, Wu H. Enterprise knowledge graph: An introduction. InExploiting linked data and knowledge graphs in large organisations 2017 (pp. 1-14). Springer, Cham. [Consultado 20 julio de 2022]. Disponible en: https:/doi.org/10.1007/978-3-319-45654-6_1.
19. Sequeda J, Lassila O. Designing and building enterprise knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge. 2021 Aug 3;11(1):1-65. [Consultado 3 agosto de 2022]. Disponible en: https://doi.org/10.2200/S01105ED1V01Y202105DSK020.
20. Cárdenas ML, Fernández TD, Fernández MD, de la Iglesia Campos M. GRAFOS VIRTUALES DE CONOCIMIENTO PARA LA INTEGRACIÓN DE DATOS EMPRESARIALES EN UNA EMPRESA CUBANA. Revista Cubana de Administración Pública y Empresarial. 2022 Apr 20;6(1):e211. [Consultado 14 julio de 2022]. Disponible en: https://doi.org/10.5281/zenodo.6472957.
21. Xiao G, Ding L, Cogrel B, Calvanese D. Virtual knowledge graphs: An overview of systems and use cases. Data Intelligence. 2019 Jun 1;1(3):201-23. [Consultado 2 agosto de 2022]. Disponible en: https://doi.org/10.1162/dint_a_00011

Conflicto de intereses

La autora declara no presentar conflictos de intereses