

recimundo

Revista Científica Mundo de la Investigación y el Conocimiento

DOI: 10.26820/recimundo/4.(3).julio.2020.242-250

URL: <http://recimundo.com/index.php/es/article/view/870>

EDITORIAL: Saberes del Conocimiento

REVISTA: RECIMUNDO

ISSN: 2588-073X

TIPO DE INVESTIGACIÓN: Artículo de Investigación

CÓDIGO UNESCO: 1203 Ciencia de Los Ordenadores;
1203.17 Informática

PAGINAS: 242-250



Virtualización de datos una alternativa moderna de integración de datos

Data virtualization a modern alternative for data integration

A virtualização de dados é uma alternativa moderna para integração de dados

Jorge Eduardo Cevallos Zhunio ¹; **Gustavo Adolfo Noboa Franco** ²;
Carlos Alex Valle Chiriboga ³; **Pamela Elizabeth Rodriguez Velasquez** ⁴

RECIBIDO: 10/04/2020 **ACEPTADO:** 26/05/2020 **PUBLICADO:** 01/07/2020

1. Magister en Sistemas de Información mención en Inteligencia de Negocios; Ingeniero en Sistemas Administrativos Computarizados; Universidad de Guayaquil; cevalloszhunio04@gmail.com;  <https://orcid.org/0000-0001-8976-2973>
2. Magister en Sistemas de Información mención en Inteligencia de Negocios; Ingeniero en Sistemas Computacionales; gustavonoboa19@hotmail.com;  <https://orcid.org/0000-0002-6907-4863>
3. Magister en Sistemas de Información mención en Inteligencia de Negocios, Licenciado en Sistemas de Información; carlitos.alex.valle@gmail.com;  <https://orcid.org/0000-0003-0090-378X>
4. Magister en Sistemas de Información mención en Inteligencia de Negocios, Ingeniera en Sistemas Computacionales; pamelizabth@hotmail.com;  <https://orcid.org/0000-0002-8371-9824>

CORRESPONDENCIA

Jorge Eduardo Cevallos Zhunio
cevalloszhunio04@gmail.com

Quito, Ecuador

RESUMEN

En este trabajo se realizó una revisión bibliográfica sobre la virtualización de datos con el objetivo de exponer y consolidar información sobre esta alternativa de integración de datos, se utilizó una investigación de tipo documental teniendo como fuente de información: libros digitales e impresos, informes, artículos científicos y de revisión. Se definió a la virtualización de datos como una tecnología moderna que permite integrar datos de diferentes fuentes, creando bases de datos o tablas lógicas (denominadas vistas), las cuales se guardan en la memoria del mainframe y que están disponibles en tiempo real para usuarios y herramientas consumidoras de datos; convirtiéndose en una alternativa de integración de datos que le ofrece a los arquitectos empresariales la oportunidad de explotar las fuentes de datos con el objetivo de facilitar la analítica de datos y la gobernanza de datos. Como toda tecnología está supeditada a la necesidad de la empresa, los costos de adquisición e implementación deben estar conforme a los beneficios y al cumplimiento de objetivos institucionales.

Esta tecnología moderna de integración de datos puede ser utilizada en proyectos con fuentes de datos heterogéneas permitiéndoles a los arquitectos empresariales enfocarse en los datos y no desgastarse en complejas y múltiples herramientas tecnológicas, además de ser un aliciente para nuevas iniciativas comerciales, representa ventajas económicas, tecnológicas y procedimentales, sin embargo, podríamos concluir que no es el reemplazo absoluto de otras tecnologías de integración, pero sí podría acoplarse fácilmente a las tecnologías que tiene actualmente la empresa.

Palabras clave: Virtualización de datos, integración de datos, ETL, analítica empresarial, gobernanza de datos.

ABSTRACT

In this work, a bibliographic review on data virtualization was carried out in order to expose and consolidate information on this data integration alternative, a documentary-type research was used having as a source of information: digital and printed books, reports, articles scientific and review. Data virtualization was defined as a modern technology that allows integrating data from different sources, creating databases or logical tables (called views), which are stored in the memory of the mainframe and are available in real time to users and data consuming tools; becoming a data integration alternative that offers enterprise architects the opportunity to exploit data sources in order to facilitate data analytics and data governance. As all technology is subject to the need of the company, the acquisition and implementation costs must be in accordance with the benefits and the fulfillment of institutional objectives.

This modern data integration technology can be used in projects with heterogeneous data sources, allowing business architects to focus on data and not wear out complex and multiple technological tools, in addition to being an incentive for new business initiatives, it represents economic advantages, technological and procedural, however, we could conclude that it is not the absolute replacement of other integration technologies, but it could easily be coupled to the technologies that the company currently has.

Keywords: Data virtualization, data integration, ETL, business analytics, data governance.

RESUMO

Neste trabalho, foi realizada uma revisão bibliográfica sobre virtualização de dados com o objetivo de expor e consolidar informações sobre esta alternativa de integração de dados, foi utilizada uma pesquisa do tipo documental tendo como fonte de informação: livros digitais e impressos, relatórios, artigos científicos e Reveja. A virtualização de dados foi definida como uma tecnologia moderna que permite integrar dados de diferentes fontes, criando bancos de dados ou tabelas lógicas (chamadas de views), que ficam armazenadas na memória do mainframe e ficam disponíveis em tempo real aos usuários e ferramentas consumidoras de dados; tornando-se uma alternativa de integração de dados que oferece aos arquitetos corporativos a oportunidade de explorar as fontes de dados para facilitar a análise e governança de dados. Como toda tecnologia está sujeita à necessidade da empresa, os custos de aquisição e implantação devem estar de acordo com os benefícios e o cumprimento dos objetivos institucionais.

Esta moderna tecnologia de integração de dados pode ser utilizada em projetos com fontes de dados heterogêneas, permitindo aos arquitetos de negócios focar nos dados e não desgastar ferramentas tecnológicas complexas e múltiplas, além de ser um incentivo para novas iniciativas de negócios, representa vantagens econômicas, tecnológicas e processual, porém, podemos concluir que não se trata da substituição absoluta de outras tecnologias de integração, mas poderia facilmente ser acoplada às tecnologias que a empresa possui atualmente.

Palavras-chave: Virtualização de dados, integração de dados, ETL, análise de negócios, governança de dados.

Introducción

Las empresas públicas y privadas sin importar su tamaño generan una gran cantidad de datos todos los días y normalmente los datos son generados por diferentes fuentes entre las que encontramos almacenes de datos, bases de datos, sistemas transaccionales, redes sociales e incluso correos. Todas estas fuentes de datos trabajan con tecnologías diferentes, sin embargo, el tenerlas integradas al alcance de los usuarios es una necesidad para la toma de decisiones.

(Gartner, 2020) define la integración de datos como una disciplina que comprende las prácticas, las técnicas arquitectónicas y las herramientas que permiten a las organizaciones ingerir, transformar, combinar y suministrar datos en todo el espectro de tipos de datos. Esta integración se lleva a cabo en la empresa y más allá, tanto en socios como en casos de uso y fuentes de datos de terceros, para cumplir con los requisitos de consumo de datos de todas las aplicaciones y procesos comerciales. Esto incluye cualquier tecnología que admita los requisitos de integración de datos independientemente de la nomenclatura actual del mercado (por ejemplo, ingesta de datos, transformación de datos, procesamiento de datos, canalización de datos, replicación de datos, sincronización de datos, virtualización de datos, integración de datos de flujo, servicios de datos, estructuras de datos, ingeniería de datos y muchos más).

La integración de datos es una serie de técnicas y procesos que se encargan de llevar a cabo todas las tareas relacionadas con la extracción, manipulación, control, integración, depuración de datos, carga y actualización del DW, es decir, todas las tareas que se realizarán desde que se obtienen los datos de los diferentes Data Sources hasta que se cargan en el DW (Bernabeu & García, 2017). La integración de datos es el proceso de combinar datos de un conjunto posiblemente heterogéneo de almacenes

de datos para crear una vista unificada de todos esos datos (Van der Lans, 2012).

Los datos son integrados con el objetivo de proporcionar información relevante a los usuarios, existen algunas tecnologías de integración de datos entre las que más se utilizan están el ETL, y el ESB; para (Bernabeu & García, 2017) el ETL (Extraction, Transformation, Load) es solo una de las muchas técnicas de la integración de datos, que incluye los procesos de extracción, transformación y carga: el proceso extracción incluirá técnicas enfocadas por ejemplo a obtener desde diversas fuentes solamente los datos relevantes y mantenerlos en una Staging Area (almacenamiento intermedio); el proceso transformación incluirá técnicas encargadas de compatibilizar formatos, filtrar y clasificar datos, relacionar diversas fuentes, etc; el proceso carga incluirá técnicas propias de la carga de datos y actualización del DW. El ETL puede ser llevado a cabo en diferentes sistemas gestores de base de datos como Microsoft SQL Server, MySQL, Oracle Database, PostgreSQL, entre otros.

Un ESB es un entorno que pertenece a la categoría de sistemas de middleware de plataforma, que proporciona una sofisticada interconectividad entre servicios y permite superar problemas relacionados con la confiabilidad, escalabilidad y comunicación. Proporciona mecanismos que permiten integrar sistemas distribuidos siguiendo el paradigma promovido por las Arquitectura Orientada a Servicios (SOA) (Llambías & Ruggia, 2015).

Una alternativa para la integración de datos es la virtualización de datos que es un enfoque moderno para la integración de datos. Trasciende las limitaciones de las técnicas tradicionales al ofrecer una vista simplificada, unificada e integrada de datos comerciales confiables en tiempo real o casi en tiempo real, según sea necesario al consumir aplicaciones, procesos, análisis o usuarios comerciales (Miller, 2019).

Metodología

Para (Pulido, 2015) la metodología es el procedimiento a través del cual se establecen una serie de normas, de obligado cumplimiento para el investigador, que buscan la validez lógica de la realidad investigada en pos de posibilitar la predicción y la explicación de cuestiones significativas, finalidad última de la investigación científica.

En este artículo se realiza una revisión bibliográfica sobre la virtualización de datos con el objetivo de exponer y consolidar información sobre esta alternativa de integración de datos, el análisis retrospectivo de la literatura disponible en fuentes digitales y la experiencia de los investigadores va a permitir dilucidar tópicos del tema en mención.

Tipo de Investigación

Las actividades y herramientas efectuadas en esta investigación conllevan a establecer que se realizó una investigación documental determinada por (Ortega, 2017) como aquella que se apoya en fuentes documentales. Como subtipos de esta investigación, se encuentra la investigación bibliográfica, la hemerográfica y la archivística; la primera se basa en la consulta de libros, la segunda en artículos o ensayos de revistas y periódicos, y la tercera en documentos que se encuentran en los archivos, como cartas, oficios, circulares, expedientes, etcétera.

Técnicas para la Recolección de la Información

En la recolección de datos se utilizó fuentes de datos secundarios (libros digitales e impresos, informes, artículos científicos y de revisión), que según explica (Pulido, 2015) en el proceso de investigación se distinguen dos tipos fundamentales de datos: los primarios y los secundarios. Los primarios hacen referencia a informaciones recogidas de primera mano, mientras que los secundarios son datos extraídos de los datos originales recogidos por otras personas.

Para el análisis y selección de la informa-

ción, proporcionados por las fuentes de datos, se utilizó fichas de registros de información, técnica del subrayado y resúmenes. Las fuentes de información de preferencia fueron: buscadores académicos, bases de datos bibliográficas, bibliotecas electrónicas, sistemas de información de revistas científicas, sitios web de proveedores de tecnología.

Resultados

La virtualización de datos no es una tecnología completamente nueva, sin embargo, a nivel nacional prácticamente es desconocida a diferencia que en otros países sus usos y beneficios son aprovechados por diferentes tipos de empresa; en el 2011 (Bologa & Bologa) en un artículo denominado "A Perspective on the Benefits of Data Virtualization Technology", se refería a la entonces inédita tecnología como: una herramienta muy importante y prometedora que se puede utilizar de forma independiente o complementaria a la integración de datos físicos. (Van der Lans, 2012) la definía como una alternativa revolucionaria de los data warehouses, sin embargo, la virtualización de datos no sólo es una alternativa a los almacenes de datos, sino, una opción moderna de integración de datos.

Actualmente un sin número de proyectos han utilizado esta tecnología como es el caso de la Universidad Nacional de Río Negro que utilizó la virtualización de datos en su proyecto "Sistema de Administración de Información de Proyectos e Investigadores - (SAIPI)", en base a lo afirmado por (Cambarieri, Hünicken, García, & Petroff, 2016) donde destacan que la virtualización de datos permitió la integración con el sistema SIGEVA ofreciendo los datos necesarios, en la forma requerida, de manera centralizada, y en tiempo real para la administración, seguimiento y gestión de los proyectos de investigación, además indican que la decisión de utilizar la capa de virtualización de datos, permitió que el equipo se centre en el análisis de los datos y no en cuestiones

técnicas relativas a la fuente de los mismos. Parecidos criterios emiten (Suárez Blanco & Plasencia Salgueiro, 2016) que utilizaron la Plataforma Denodo para la virtualización de datos del Sistema de Vigilancia Tecnológica e Investigación Beagle, los autores afirman que la virtualización de datos proporciona información abstracta e integrada en tiempo real a partir de fuentes dispares a múltiples aplicaciones y usuarios. Pero también es fácil de construir, fácil de consumir y mantener.

La virtualización de datos para (Borden, Fudge, Nelson, & Porell, 2018) es un método de integración de datos que reúne fuentes de datos dispares para crear vistas integradas y virtualizadas de los datos. Estas vistas o tablas virtuales se guardan en la memoria y están aptas para las aplicaciones como fuentes de datos lógicas.

La virtualización de datos es una tecnología moderna que permite integrar datos de diferentes fuentes, creando bases de datos o tablas lógicas (denominadas vistas), las cuales se guardan en la memoria del mainframe y que están disponibles en tiempo real para usuarios y herramientas consumidoras de datos. Es una tecnología que podría facilitar a las empresas que buscan

implementar una gobernanza de datos definido por (Soares, 2014) como la formulación de políticas para optimizar, proteger y aprovechar la información como un activo empresarial alineando los objetivos de múltiples funciones.

Esta tecnología ofrece la capacidad de aprovechar grandes volúmenes de datos de IoT (internet de las cosas) de sensores y dispositivos almacenados en repositorios Hadoop, Spark o NoSQL, junto con datos históricos para realizar análisis (Forrester, 2017). Incluso las empresas podrían usar las herramientas de virtualización de datos como un soporte hacia la analítica empresarial que según (Laursen & Thorlund, 2016) es una disciplina holística y jerárquica, que se extiende desde las estrategias comerciales hasta la obtención de fuentes de datos operativos. El entorno impulsado por la empresa debe asumir la plena propiedad y gestionar el proceso. El entorno técnicamente orientado debe respaldar el proceso con infraestructura, entrega de datos y la funcionalidad de aplicación necesaria.

En la siguiente figura se muestra un esquema de virtualización de datos conformado esencialmente por: fuentes de datos, virtualización de datos y consumidores de datos.

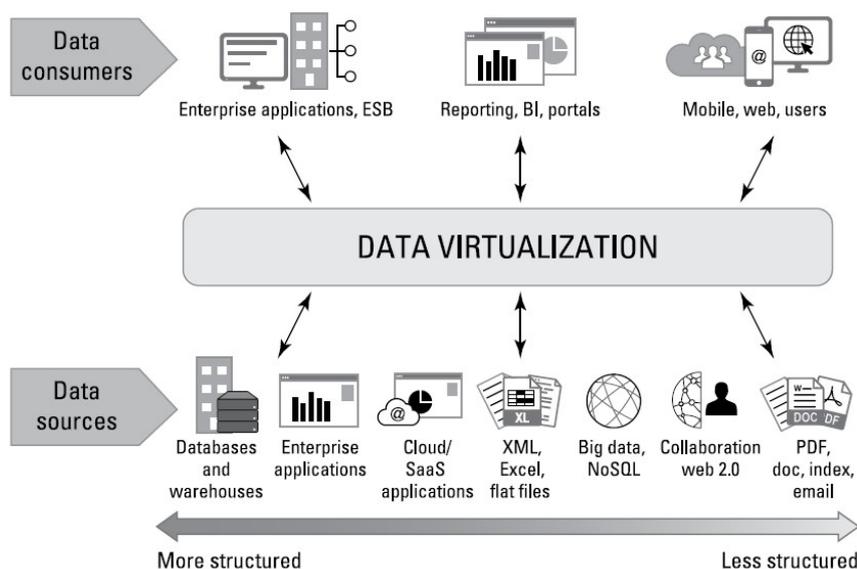


Figura 1. Esquema de virtualización de datos

Data sources traducidas al español como fuente de datos son todas esas herramientas tecnológicas que generan datos entre las que tenemos a los almacenes de datos (data mart, data warehouse), bases de datos, data lakes, clústeres de Hadoop, aplicaciones empresariales, datos en la nube, archivos en general e incluso correos.

Virtualización de datos es la capa intermedia entre las fuentes de datos y los consumidores de datos, funciona como una capa de abstracción y una capa de servicios de datos. En la actualidad los proveedores de virtualización tienen su propia GUI (interfaz gráfica de usuario) donde el usuario puede acceder a los datos integrados y realizar consultas (Query).

Data consumers traducidas al español como consumidores de datos son las herramientas tecnológicas encargadas de recibir los datos con el objetivo de presentarlas a un usuario final, entre las que destacan las herramientas de inteligencia de negocios, páginas web, aplicaciones empresariales o incluso aplicaciones móviles.

¿Virtualización de datos reemplazo o alternativa de tecnologías tradicionales de integración de datos?

La construcción de un ETL (extracción, transformación y carga) ha sido la vía tradicional para la integración de los datos y su almacenamiento, sin embargo, este proceso es largo y alto consumidor de recursos (Suárez Blanco & Plasencia Salgueiro, 2016).

La virtualización de datos realiza muchas de las mismas funciones de transformación y calidad que la integración de datos tradicional, como ETL, replicación de datos, federación de datos, Enterprise Service Bus (ESB) y otros, pero aprovecha la tecnología moderna para ofrecer integración de datos en tiempo real a un costo menor, con más velocidad y agilidad. Puede reemplazar la integración de datos tradicional y reducir la necesidad de almacenes de datos y merca-

dos de datos replicados en muchos casos (Miller, 2019).

La virtualización de datos puede ser un complemento estratégico para los almacenes de datos o lagos de datos al proporcionar a las organizaciones opciones, como el acceso a datos virtuales, que deja los datos en su lugar y, en esencia, crea un almacén de datos lógico (Borden, Fudge, Nelson, & Porell, 2018)

A medida que aumenta el tamaño de los datos y su heterogeneidad, la arquitectura tradicional del sistema de bases de datos se convierte en un obstáculo para el análisis de datos. Las consultas, por otro lado, a menudo son ad-hoc y están respaldadas por operadores precocinados que no son lo suficientemente adaptables para optimizar el acceso a los datos (Karpathiotakis, Alagiannis, Heinis, & Branco, 2015).

El software de virtualización de datos se implementa rápidamente, se adapta fácilmente al cambio (agilidad) y puede entregar simultáneamente diferentes datos para muchos usos diferentes a muchos consumidores de datos (Ferguson, 2014). A pesar de esto para (Gartner, 2020) más del 50% de las organizaciones ahora esperan que las herramientas de integración de datos brinden un estilo de interoperabilidad de integración, es decir que los software de integración de datos sean capaces de combinar y cambiar entre ETL, replicación y virtualización según las demandas de sus casos de uso.

Proveedores y herramientas para virtualizar datos

Existen varios proveedores de virtualización de datos entre los que según (Forrester, 2017) destacan: Informatica, Denodo Technologies, IBM, SAP, TIBCO Software. Los proveedores fueron evaluados en base a:

Las características arquitectónicas y operativas y la funcionalidad de cada solución; cómo planea evolucionar sus servicios de virtualización de datos empresariales, el

compromiso y las estrategias de dirección de comercialización de cada proveedor; las finanzas, la adopción y las asociaciones de las empresas de cada proveedor.

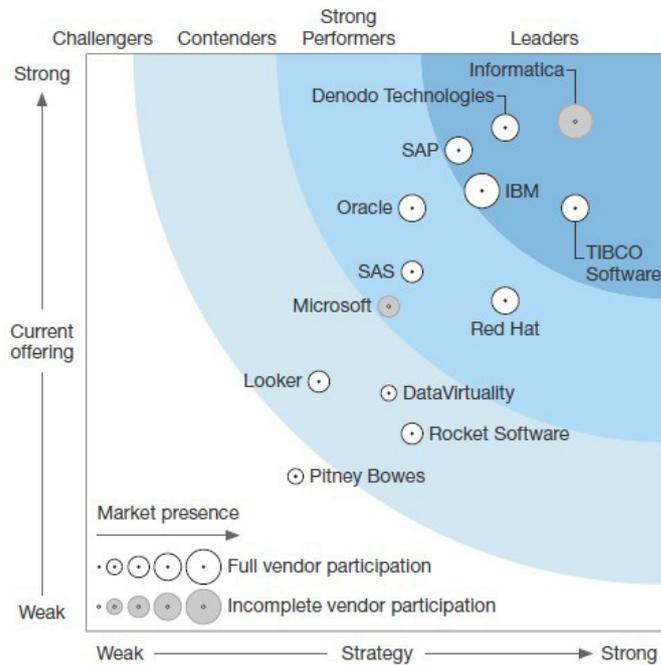


Figura 2. Forrester Wave™: Enterprise Data Virtualization, Q4 2017

La virtualización de datos es una excelente alternativa de integración de datos y esto se ratifica en el informe Magic Quadrant for Data Integration Tools realizado por (Gartner, 2020), en el que destacan varios proveedores de virtualización de datos como herramientas de integración de datos.



Figura 3. Magic Quadrant for Data Integration Tools

Conclusiones

La virtualización de datos es una alternativa moderna de integración de datos con ventajas económicas, tecnológicas y procedimentales, sin embargo, podríamos concluir que no es el reemplazo absoluto de otras tecnologías de integración, pero sí podría acoplarse fácilmente a las tecnologías que tiene actualmente la empresa.

La virtualización de datos da libertad para realizar consultas, a diferencia de otras tecnologías de integración, es ideal usarla en proyectos con fuentes de datos heterogéneas permitiéndole a los arquitectos empresariales enfocarse en los datos y no desgastarse en complejas y múltiples herramientas tecnológicas, además de ser un aliciente para nuevas iniciativas comerciales, ya que permite al usuario final manipular los datos de forma dinámica y en tiempo real.

En primera instancia la virtualización de datos era una alternativa a los almacenes de datos (Data Warehouse, Data Mart), sin embargo, actualmente la virtualización de datos es una alternativa de integración de datos que le ofrece a los arquitectos empresariales la oportunidad de explotar las fuentes de datos con el objetivo de facilitar la analítica de datos y la gobernanza de datos. Como toda tecnología está supeditada a la necesidad de la empresa, los costos de adquisición e implementación deben estar conforme a los beneficios y al cumplimiento de objetivos institucionales.

Bibliografía

- Bernabeu, D., & García, M. (2017). Hefesto Data Warehousing, V3. Córdoba.
- Bologa, A., & Bologa, R. (2011). A Perspective on the Benefits of Data Virtualization Technology. *Informática Eómica* vol. 15, 110-118.
- Borden, B., Fudge, C., Nelson, J., & Porell, J. (2018). *Accelerating Digital Transformation on Z Using Data Virtualization*. IBM RedBooks.
- Cambarieri, M., Hünicken, L., García, N., & Petroff, M. (2016). Virtualización de datos: una solución para la integración de datos y extensión de funcionalidades de sistemas existentes. 45 JAIIO, 157-169.
- Ferguson, M. (2014). Data virtualization-Flexible Technology for the Agile Enterprise Intelligent Business Strategies. *Intelligent Business Strategies*, 3-20.
- Forrester. (15 de Noviembre de 2017). Report. Obtenido de Forrester: <https://www.forrester.com/report/The+Forrester+Wave+Enterprise+Data+Virtualization+Q4+2017/-/E-RES133042>
- Gartner. (18 de Agosto de 2020). Obtenido de Gartner Web site: <https://www.gartner.com/doc/reprints?id=1-1ZDKIKNH&ct=200701&st=sb>
- Karpathiotakis, M., Alagiannis, I., Heinis, T., & Branco, M. (2015). Just-In-Time Data Virtualization: Lightweight Data Management with ViDa. *PVLDB*, 43-59.
- Laursen, G., & Thorlund, J. (2016). *Business Analytics for Managers : Taking Business Intelligence Beyond Reporting*. New Jersey: John Wiley & Sons.
- Llambías, G., & Ruggia, R. (2015). A middleware-based platform for the integration of bioinformatic services. *CLElej*.
- Miller, L. (2019). *Data Virtualization For Dummies*. West Sussex: John Wiley & Sons, Ltd.
- Ortega, G. (2017). Cómo se genera una investigación científica que luego sea motivo de publicación. *Journal of the Selva Andina Research Society*, 145-146.
- Pulido, M. (2015). Ceremonial y protocolo: métodos y técnicas de investigación científica. *Opción*, 1137-1156.
- Soares, S. (2014). *Data Governance Tools : Evaluation Criteria, Big Data Governance, and Alignment with Enterprise Data Management*. Mc Press Online.
- Suárez Blanco, I., & Plasencia Salgueiro, A. (2016). Plataforma Denodo para la virtualización de datos del Sistema de Vigilancia Tecnológica e Investigación Beagle. La Habana, Cuba. Obtenido de <http://www.congreso-info.cu/index.php/info/2016/paper/viewFile/304/323>
- Van der Lans, R. (2012). *ata Virtualization for Business Intelligence Systems : Revolutionizing Data Integration for Data Warehouses*. Waltham: Elsevier Science & Technology.



RECONOCIMIENTO-NOCOMERCIAL-COMPARTIRIGUAL
CC BY-NC-SA

ESTA LICENCIA PERMITE A OTROS ENTREMEXCLAR, AJUSTAR Y
CONSTRUIR A PARTIR DE SU OBRA CON FINES NO COMERCIALES, SIEMPRE
Y CUANDO LE RECONOZCAN LA AUTORÍA Y SUS NUEVAS CREACIONES
ESTÉN BAJO UNA LICENCIA CON LOS MISMOS TÉRMINOS.

CITAR ESTE ARTICULO:

Cevallos Zhunio, J., Noboa Franco, G., Valle Chiriboga, C., & Rodriguez Velasquez, P. (2020). Virtualización de datos una alternativa moderna de integración de datos. RECIMUNDO, 4(3), 242-250. doi:10.26820/recumundo/4.(3). julio.2020.242-250