DATA LAKES: APLICACIONES, HERRAMIENTAS Y ARQUITECTURAS

JAVIER CAMILO AGUDELO PATIÑO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
PEREIRA
2020

DATA LAKES: APLICACIONES, HERRAMIENTAS Y ARQUITECTURAS

JAVIER CAMILO AGUDELO PATIÑO

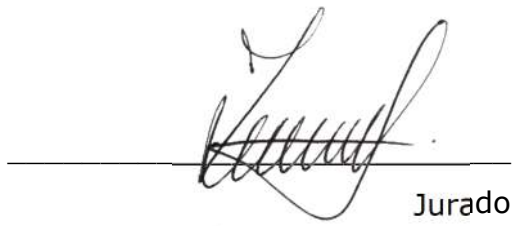
Monografía presentada como requisito para optar al Título de Ingeniero de Sistemas y Computación

Directora del Trabajo de Grado
Ingeniera Ivonne Castaño Osorio

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
PEREIRA
2020

NOTA DE ACEPTACIÓN

APROBADO



Jurado

Pereira, Julio 29 de 2020

DEDICATORIA

Este proyecto esta dedicado a mi familia, mis padres por haberme forjado como persona íntegra, con un apoyo, ejemplo y motivación continua para alcanzar mis metas, a mis amistades más cercanas que han sido un gran ejemplo y han fomentado en mi un deseo de superación y triunfo en la vida, lo que ha contribuido a la consecución de este logro, a mi directora de proyecto de grado Ivonne Castaño Osorio, que sin su ayuda y conocimiento no hubiese sido posible realizar este proyecto, a todas esas personas que ayudaron para que el mismo pudiera culminarse y que han aportado enseñanzas y aprendizajes que cada día hacen de mí una mejor persona.

CONTENIDO

INTRODUCCIÓN

La aparición de la Big Data ha demostrado su utilidad como la herramienta más importante que las empresas ponen en práctica para modelar su futuro. Empresas de gran envergadura utilizan los Big Data para introducir a una velocidad vertiginosa su innovación en todos los ámbitos, desde el compromiso con los clientes al desarrollo de productos nuevos o la estrategia de optimización de la empresa. El auge de las tecnologías de Big Data, combinado con la promesa de la tecnología Cloud, ha facultado a innumerables empresas para implantar sus iniciativas de Big Data sin ningún esfuerzo. Avanzando hacia el universo Cloud, las empresas ya están recogiendo frutos, como la velocidad de aprovisionamiento, el tiempo de llegada al mercado, la flexibilidad y agilidad, la escalabilidad instantánea o la reducción de los gastos generales de informática y empresa, por poner tan solo algunos ejemplos. Sin embargo, los repositorios de datos muchas veces no están a la altura de las necesidades de las organizaciones, quienes, a pesar de contar con avanzadas herramientas de Big Data, no pueden sacarles el máximo provecho.

Debido a esto, los Data Lake son un enfoque arquitectónico emergente y poderoso, especialmente a medida que las empresas recurren a aplicaciones móviles basadas en la nube e Internet of Things (IoT) como medios de entrega en tiempo adecuado para Big Data, pero como enfoque emergente, aún hay muchas dudas acerca de su verdadera funcionalidad y las ventajas que supone su aplicación.

Esta monografía pretende dar a conocer en qué estado de desarrollo se encuentran los data lakes, sus características, en qué campos se están aplicando actualmente y las arquitecturas que se han venido trabajando para dichas aplicaciones.

1. PLANTEAMIENTO DEL PROBLEMA

En la era de los grandes datos, los teléfonos inteligentes, las redes sociales, los objetos conectados y otros creadores de datos, se genera un gran volumen de datos estructurados, semiestructurados y no estructurados mucho más rápido que antes. Estos datos tienen un gran valor para los Sistemas de Soporte de Decisiones de las organizaciones, cuya piedra angular se basa en los datos. Sin embargo, manejar datos heterogéneos y voluminosos es especialmente desafiante para este tipo de sistemas de información.

Hoy en día, el Data Warehouse es una solución de uso común en los sistemas de soporte de decisiones DSS. Los datos se han extraído, transformado y cargado mediante procesos ETL, de acuerdo con esquemas predefinidos. Data Warehouse es popular gracias a su respuesta rápida, rendimiento constante y análisis de funciones cruzadas. Sin embargo, no están adaptados para el análisis de Big Data porque sólo se pueden responder requisitos predefinidos, parte de la información se pierde a través de los procesos ETL, y el costo de un DW puede crecer exponencialmente debido a los requisitos de un mejor rendimiento, el crecimiento del volumen de datos y la complejidad de la base de datos.

Para enfrentar los desafíos de los grandes datos y las deficiencias de DW, se presentó el concepto de lago de datos o Data Lake (DL): "Si un almacén de datos es como agua embotellada, limpia y estructurada para un consumo fácil de los datos, el Data Lake es un gran cuerpo de agua en un estado más natural"[1]. Esta explicación esboza el esquema de DL, pero no puede considerarse como una definición formal. Data Lake es un concepto relativamente nuevo, y como tal, no tiene una definición estándar ni una arquitectura reconocida, por lo que se hace importante analizar sus diferentes definiciones, describir sus características generales, revisar las diversas arquitecturas que se han planteado para las diferentes aplicaciones, y conocer las principales herramientas que han

[1] James Dixon. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes

permitido su implementación en las organizaciones, para así dar una visión más central y cercana a lo que en realidad se configura como Data Lake.

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Describir qué son los Data Lake, sus características, aplicaciones prácticas, herramientas mediante las cuáles se están construyendo y las diferentes arquitecturas utilizadas.

2.2. OBJETIVOS ESPECÍFICOS

- ✓ Definir qué es un Data Lake.
- ✓ Describir las arquitecturas de los Data Lakes.
- ✓ Realizar análisis comparativo de las arquitecturas descritas.
- ✓ Identificar las herramientas utilizadas en el entorno de los Data Lakes.
- ✓ Identificar las diferentes aplicaciones que se le han dado a los Data Lakes.

3. JUSTIFICACIÓN

Esta monografía permitirá facilitar la compresión de la tecnología de los lagos de datos, pues por ser un fenómeno reciente, la documentación disponible se centra en algunos detalles específicos de los lagos de datos, pero no recogen en un solo documento las diferentes arquitecturas y sus enfoques. Este documento pretende ser una guía de estudio que permita entender qué son, de dónde provienen, cómo están conformados, cuáles son sus puntos clave y qué herramientas se utilizan en ellos.

4. METODOLOGÍA

Esta investigación tiene un enfoque cualitativo ya que se pretende realizar una monografía descriptiva del concepto de Data Lake y los principales elementos que le conciernen.

Las fases para el desarrollo de la monografía serán:

- ✓ Definición de las fuentes a consultar.
- ✓ Recopilación de información sobre los Data Lakes.
- ✓ Clasificación y análisis de la información recopilada.
- ✓ Construcción del documento.

Las fuentes primarias serán bases de datos especializadas, revistas científicas, libros sobre el tema, y redes colaborativas de investigadores tales como Research Gate.

5. ¿QUÉ ES UN DATA LAKE?

Big data es una cantidad de datos con la que no se puede lidiar con el uso de los métodos tradicionales. Se refiere a un conjunto de datos que se extiende más allá de la capacidad de los sistemas disponibles para almacenar y manipular. Son volúmenes dinámicos, grandes y dispares de datos creados por personas, herramientas y máquinas.

Para poder trabajar ordenadamente, se definió un gobierno de Big Data que requiere tres cosas:

- Integración automatizada, es decir, fácil acceso a los datos donde sea que resida
- Contenido visual, es decir, categorización fácil, indexación y descubrimiento dentro de Big Data para optimizar su uso
- Gobernanza ágil, definición y ejecución de gobernanza adecuada al valor de los datos y su uso previsto.

Big Data requiere una tecnología escalable para recolectar, alojar y procesar analíticamente gran cantidad de datos, y para este propósito se creó Hadoop, el cual es un marco de software de código abierto desarrollado para gestionar la Big Data. Se implementa en varios módulos especializados distintos: Almacenamiento, que emplea principalmente el Sistema de archivos Hadoop, o HDFS, gestión de recursos y programación para tareas computacionales, modelos de programación de procesamiento distribuido basados en MapReduce, utilidades comunes y bibliotecas de software necesario para toda la plataforma Hadoop.

En poco tiempo, Big Data redefinió nuestra propia concepción de los datos. El gran volumen de datos que podrían almacenarse y analizarse con los sistemas Big Data revolucionó no sólo la industria, sino también el mundo permitiendo volúmenes de almacenamiento ilimitados.

Esta nueva manera de ver los datos tiene un conjunto de atributos que le definen:

- El 80-90% de las empresas manejan datos estructurados o semiestructurados (no desestructurados).
- La fuente de los datos suele ser una sola aplicación o sistema.
- Los datos son típicamente subtransaccionales o no transaccionales.
- Hay algunas preguntas conocidas para hacer de los datos.
- Hay muchas preguntas desconocidas que surgirán en el futuro.
- Hay diversas comunidades de usuarios que tienen preguntas sobre los datos.
- Los datos son de una escala o volumen diario de tal manera que no encajan técnica y/o económicamente en un sistema de gestión de bases de datos relacionales.

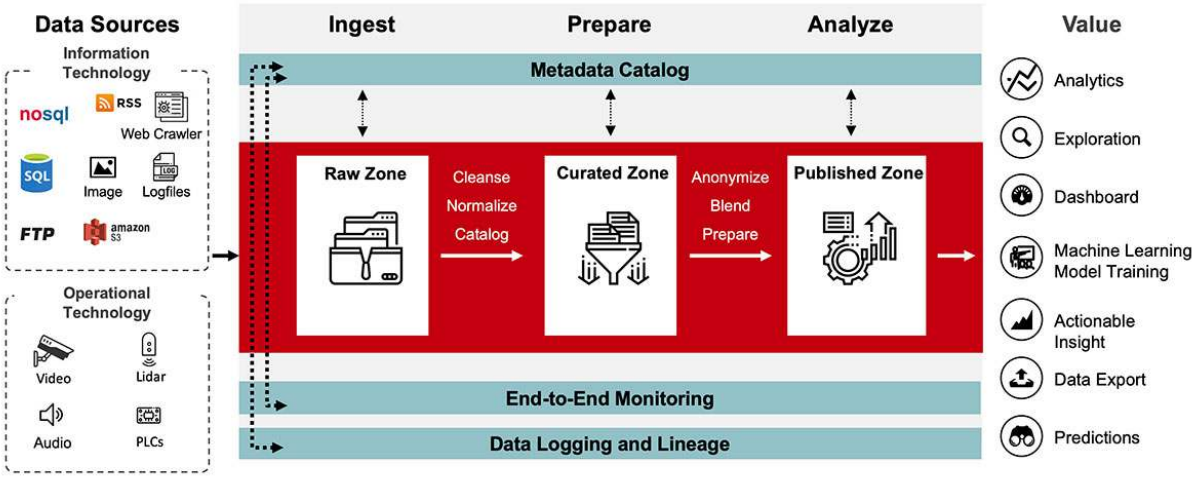
La forma estándar de manejar la presentación de informes y el análisis de estos datos era identificar los atributos más interesantes y agregarlos en un mercado de datos, pero comenzaron a presentarse varios problemas con este enfoque:

- Solo se examina un subconjunto de los atributos, por lo que solo se pueden responder preguntas predeterminadas.
- Los datos se agregan para que se pierda visibilidad en los niveles más bajos.

Teniendo presente los atributos anteriores y los problemas de las soluciones tradicionales, James Dixon, entonces CTO de Pentaho, una compañía de software de inteligencia de negocios, acuñó el concepto de Data Lake o Lago de Datos en 2010, que permite describir una solución óptima.

Un lago de datos es un repositorio o sistema de almacenamiento y análisis escalable para grandes cantidades de datos de cualquier tipo, retenidos en su formato nativo y utilizado principalmente por especialistas en datos (estadísticos, científicos de datos o analistas) para la extracción de conocimiento mediante tecnologías de bajo costo que mejoran la captura, refinamiento, archivo y exploración de datos sin procesar dentro de una empresa. Un lago de datos contiene un conjunto de datos desordenados que pueden ser estructurados, semiestructurados o no estructurados, e incluyen tablas, archivos de texto, registros del sistema y más. Sus características incluyen un catálogo de metadatos que refuerza la calidad de los datos, políticas y herramientas de gobernanza de datos,

accesibilidad a varios tipos de usuarios, integración de cualquier tipo de datos, una organización lógica y física, y escalabilidad.



El lago de datos admite las siguientes capacidades [1]:

- Capturar y almacenar datos sin procesar a escala a un bajo costo: debido a que el volumen de datos continúa creciendo bruscamente, el costo del almacenamiento de datos se hizo más importante que antes.

- Almacenar muchos tipos de datos en el mismo repositorio: hay datos estructurados del DBMS[2] tradicional, datos multiestructurados que incluyen múltiples atributos que son indefinidos, y datos multimedia. Estos diferentes tipos de datos deben procesarse de diferentes maneras.

- Realizar transformaciones en los datos: El caso de uso clave para el lago de datos es realizar el procesamiento previo y la transformación ETL[3] de los datos para su posterior exploración por otro sistema.

[2] Sistema manejador de bases de datos (SGBD) o DataBase Management System (DBMS) es una colección de software muy específico, orientado al manejo de base de datos, cuya función es servir de interfaz entre la base de datos, el usuario y las distintas aplicaciones utilizadas.

[3] Proceso ETL: Extract, Transform and Load, o extracción, transformación y carga en español, es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

- Definir la estructura de los datos en el momento en que se utilizan, denominado esquema de lectura: el lago de datos evita el esfuerzo complejo y costoso de modelado e integración de datos.

- Realizar nuevos tipos de procesamiento de datos: el lago de datos debe admitir todos los datos y todas las formas de procesamiento de datos.

- Realizar análisis de un solo sujeto basados en casos de uso muy específicos: los valores de los datos en el lago de datos no están claros, por lo que las personas tienen que desarrollar análisis específicos para descubrir cómo usar estos datos.

DATA WAREHOUSES VS. DATA LAKES

El concepto de data warehouse o almacén de datos se originó en 1988 con el trabajo de los investigadores de IBM, Barry Devlin y Paul Murphy, aunque el término data warehouse fue acuñado por William H. Inmon, el cual es conocido como el padre de Data Warehousing. Inmon describió un almacén de datos como una colección de datos orientada a un tema específico, integrado, variante en el tiempo y no volátil, que soporta el proceso de toma de decisiones.

Un almacén de datos es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso. Normalmente, un almacén de datos se aloja en un servidor corporativo o en la nube. Los datos de diferentes aplicaciones de procesamiento de transacciones online (OLTP) y otras fuentes se extraen selectivamente para su uso por aplicaciones analíticas y de consultas por usuarios.

Un almacén de datos es una arquitectura de almacenamiento diseñada para datos estructurados. Durante el desarrollo de un almacén de datos se gasta una cantidad considerable de tiempo analizando las fuentes, entendiendo los procesos de negocio y perfilando los datos. Como resultado, se obtiene un modelo de datos altamente estructurado que ya está listo para la generación de informes. Una gran parte de este proceso incluye también un procedimiento de toma de decisiones. Qué datos se incluyen

y cuáles no. Por lo general, si los datos no se utilizan para responder a preguntas específicas o no son imprescindibles en algunos informes pueden excluirse. De esta manera, se simplifica el modelo y se conserva el espacio.

Los almacenes de datos generalmente consisten en información extraída de sistemas transaccionales y, por lo tanto, incluyen métricas cuantitativas así como otros atributos que los describen. Las fuentes de datos no tradicionales como pueden ser los registros del servidor web, los datos de los sensores, la actividad de una determinada red social, imágenes o los textos suelen ignorarse.

ESTRUCTURAS DE UN DATA WAREHOUSE

La arquitectura de un almacén de datos puede ser dividida en tres estructuras simplificadas: básica, básica con un área de ensayo, y básica con área de ensayo y data marts.

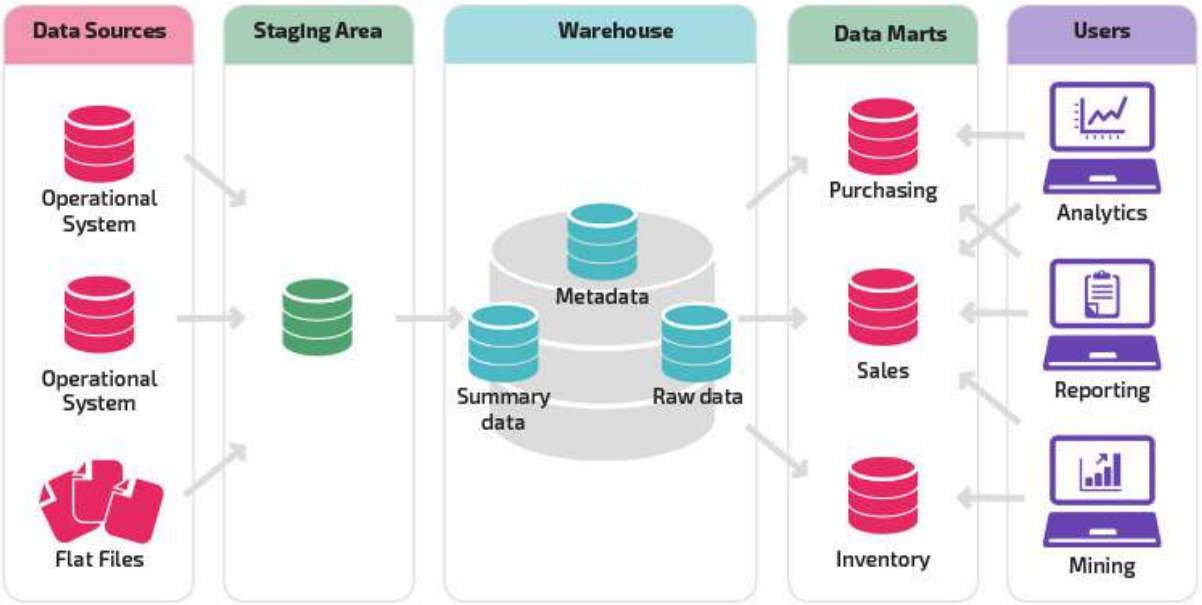
- Con una estructura básica, sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con metadatos. Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.

- Al añadir un área de ensayo que se puede colocar entre las fuentes de datos y el almacén, ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización.

- Además del área de ensayo, se pueden agregar data marts. Se pueden tener data marts separados para ventas, inventario y compras, por ejemplo, y los usuarios finales pueden acceder a datos de uno o de todos los data marts del departamento.

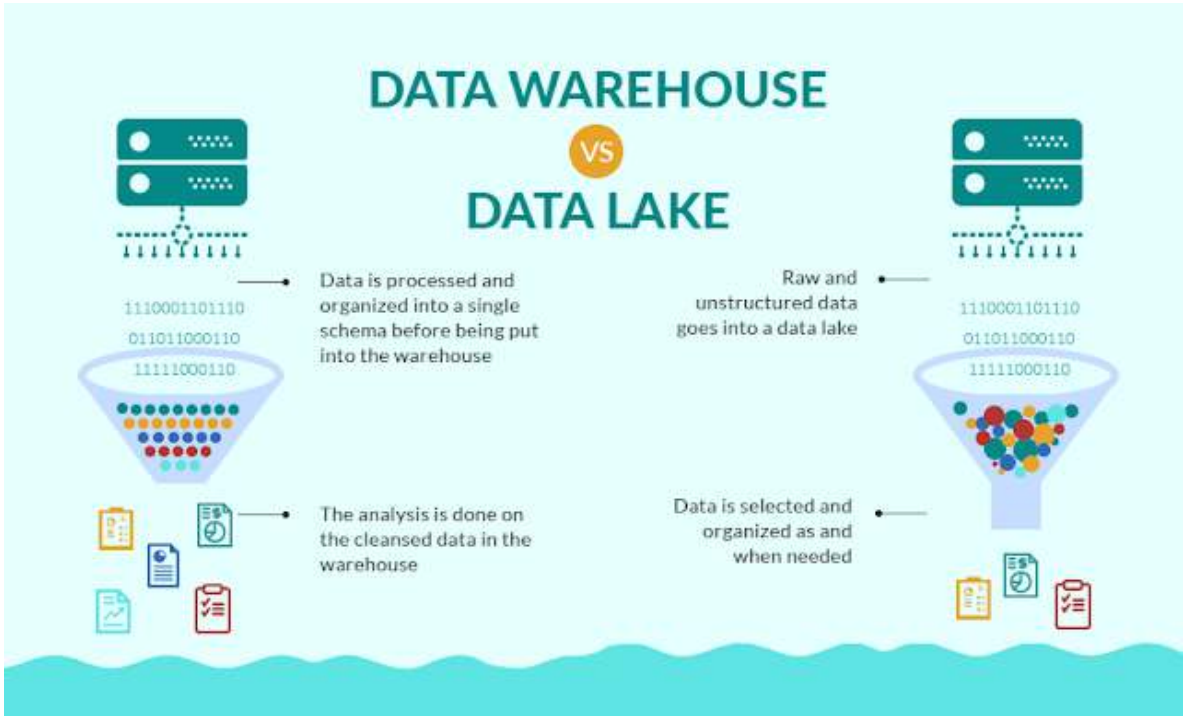
Los almacenes de datos están creados para actuar de depósito central de datos para toda una empresa, mientras que un data mart es una implementación con alcance restringido a un área funcional, problema en particular, departamento, tema o grupo de necesidades. El principal objetivo de un data mart es aislar un conjunto más pequeño

de datos del conjunto total para ofrecer un acceso más fácil a los datos para los consumidores finales.



A diferencia del almacén de datos, un lago de datos conserva todos los datos. No solo los que son vitales en ese preciso momento sino todos aquellos que están guardados que pueden hacer falta en algún momento. Se guarda todo sin tener en cuenta su estructura o su fuente. Se mantiene la información en bruto y solo se transforma en el momento en el que sea necesario, lo que proporciona varios beneficios sobre todo en cuanto a la parte de análisis. Este enfoque es posible porque el hardware para un lago de datos suele ser muy diferente del que se utiliza para un almacenamiento de datos corriente. La comodidad, los servidores que hay disponibles y el almacenamiento más barato suponen una ampliación de un lago de datos a terabytes y petabytes bastante económico.

Este enfoque se conoce como *"Schema on Read"* en comparación con el enfoque *"Schema on Write"* utilizado en el almacén de datos.



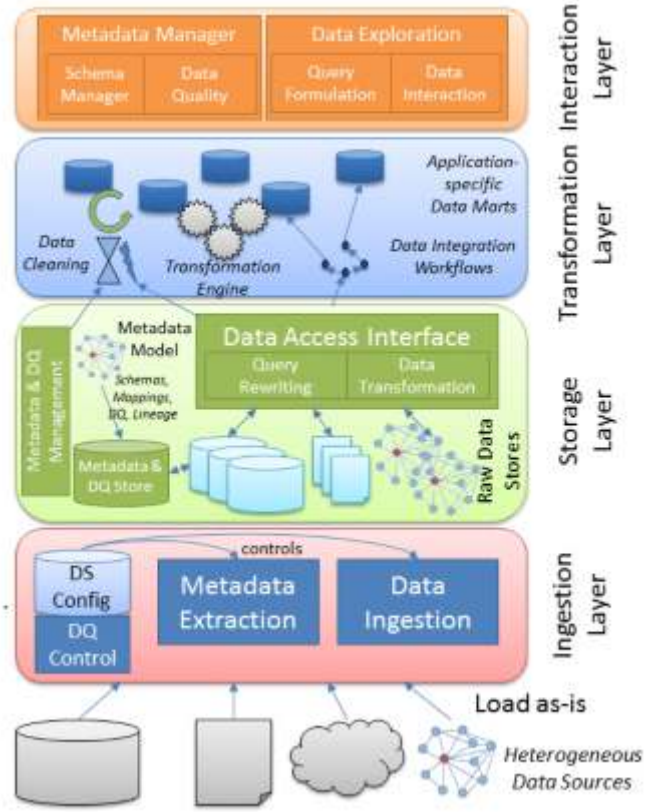
A continuación se hace un cuadro comparativo de las dos tecnologías [2]:

DIMENSIÓN	DATA WAREHOUSE	DATA LAKE
Carga de trabajo	Cientos de miles de usuarios concurrentes.	Procesamiento por lotes de datos a escala
	Realizar analíticas interactivas	Actualmente mejora sus capacidades para
	Capacidades avanzadas de gestión de la carga de trabajo.	apoyar a usuarios más interactivos
	Procesamiento por lotes	
Esquema	Normalmente, el esquema se define antes de almacenar los datos	Normalmente, el esquema se define después de almacenar los datos
	Requiere trabajo al comienzo del proceso, pero ofrece rendimiento, seguridad e integración.	Ofrece agilidad extrema y facilidad de captura de datos, pero requiere trabajo al final del proceso.
		Funciona bien para tipos de datos donde no se conoce el valor de los datos

DIMENSIÓN	DATA WAREHOUSE	DATA LAKE
Escala	Grandes volúmenes de datos a un costo moderado.	Volúmenes de datos extremos a bajo costo
Acceso	Datos a los que se accede mediante SQL estándar y herramientas de BI estandarizadas	Datos a los que se accede a través de programas creados por desarrolladores
	Método de búsqueda	Método de escaneo
Beneficio	Respuesta rápida	Excelente escalabilidad
	Rendimiento consistente	Soporte de programación
	Fácil de usar	Cambiar radicalmente
	Integración de datos	
	Análisis funcional cruzado	
	Cargue una vez, use muchos	
Consulta	SQL	Programación
Datos	Limpiado	Crudo
Costo	Uso eficiente de CPU / IO	Bajo costo de almacenamiento y procesamiento
Complejidad	Uniones complejas	Procesamiento complejo

6. ARQUITECTURA DE LOS DATA LAKES

La arquitectura funcional del lago de datos ha evolucionado de una mono zona a una zona múltiple. La primera visión de la arquitectura de un lago de datos es una arquitectura plana con una mono zona que almacena todos los datos en bruto en su formato nativo. Esta arquitectura, estrechamente vinculada al entorno HADOOP, permite cargar datos heterogéneos y voluminosos a bajo costo. Sin embargo, a pesar de todas las prestaciones que debería suponer tener un lago de datos, las organizaciones que los han implementado se han encontrado con una dificultad: están construyendo lagos de datos de "una vía", en donde los datos ingresan al lago de datos, pero nunca sale nada. Cuando esto sucede, el lago de datos pierde por completo su utilidad, convirtiéndose en un gasto para la organización, y no en la inversión que debería ser. Para remediar esta situación, Bill Inmon, ha propuesto una arquitectura que se ha convertido en la norma para construir lagos de datos eficientes [3]. Para este documento se tomará esta arquitectura como base.



6.1. ELEMENTOS BÁSICOS DE LA ARQUITECTURA

Lo primero que se plantea en esta arquitectura es que los lagos de datos eficientes necesitan cuatro elementos básicos: metadatos, mapeo de integración, contexto y metaproceso.

METADATOS

Los metadatos son la descripción de los datos en el lago de datos en sí. Los metadatos son la información estructural básica que cada colección de datos tiene asociada. Las formas típicas de metadatos incluyen descripciones del registro, los atributos, las claves, los índices y las relaciones entre los diferentes atributos de los datos. Sin embargo, hay muchas formas adicionales de metadatos. El analista utiliza los metadatos para descifrar los datos sin procesar encontrados en el lago de datos, o, en otras palabras, los metadatos son la hoja de ruta básica de los datos que residen en el lago de datos.

Cuando solo se almacenan datos sin procesar en el lago de datos, el analista que necesita usar esos datos no puede hacer nada. Los datos sin procesar por sí solos no son muy útiles. Ahora, cuando los datos sin procesar se etiquetan correctamente con metadatos y se almacenan juntos en el lago de datos, ahora tiene un servicio increíblemente útil. Este elemento se tratará a fondo en el punto 6.5.

MAPEO DE INTEGRACIÓN

El mapa de integración describe cómo los datos de una aplicación se relacionan con los datos de otra aplicación y cómo esos datos pueden combinarse significativamente. Por importantes que sean los metadatos, no es el único ingrediente básico de la arquitectura del lago de datos.

Se debe tener presente que la mayor parte de la entrada del lago de datos la genera una aplicación, de una forma u otra, así que cuando se colocan datos de diferentes aplicaciones en el lago de datos, se crean "silos" de datos no integrados en el lago de datos. Cada aplicación, generalmente escrita en un lenguaje de codificación diferente,

envía su entrada a un silo separado, que no puede comunicarse o "hablar" con los otros silos. Si bien toda la información se almacena en el mismo lago de datos, cada silo no puede integrar sus datos con los demás, incluso si están etiquetados correctamente con metadatos.

Para dar sentido a los datos en el lago de datos, es necesario crear un "mapa de integración". El mapa de integración es una especificación detallada que muestra cómo se pueden integrar los datos del lago. El mapa de integración es el mejor método para superar el aislamiento de datos en los silos.

CONTEXTO

Otro factor de complicación en el lago de datos son los datos textuales que se han colocado allí sin que se identifique el contexto del texto. El texto sin contexto son datos sin sentido. De hecho, en algunos casos es peligroso almacenar texto sin comprender su contexto. Si se va a colocar texto en el lago de datos, también debe insertar el contexto, o al menos una forma de encontrar ese contexto.

METAPROCESO

La información del metaproceso es información sobre cómo se procesaron los datos o cómo se procesará la información en el lago de datos. ¿Cuándo se generaron los datos? ¿Dónde se generaron los datos? ¿Cuántos datos se generaron? ¿Quién generó los datos? ¿Cómo se seleccionaron los datos para colocarlos en el lago de datos? Una vez dentro del lago de datos, ¿se procesaron más los datos?

Todas estas formas de meta procesamiento son útiles para el analista a medida que extrae y analiza los datos del lago. El punto más importante es que estas características deben incluirse desde el principio. Por lo general, una vez que los datos sin procesar se han cargado en el lago de datos, es demasiado tarde para volver e incluir estos ingredientes esenciales. Sin embargo, una vez que se han agregado los ingredientes, el lago de datos adquiere valor significativo.

6.2. CATEGORÍAS DE LOS DATOS EN EL LAGO DE DATOS

Si bien es cierto que se puede encontrar cualquier tipo de datos dentro del lago de datos, es posible clasificar los datos en tres categorías:

- Datos análogos
- Datos de aplicación
- Datos textuales

DATOS ANALÓGICOS

El primer tipo de datos que se encuentran en el lago de datos son los datos analógicos. Los datos analógicos generalmente son generados por una máquina o algún otro dispositivo automatizado, incluso si no están conectados a Internet. Estas herramientas de medición incluyen programas de diagnóstico que registran el rendimiento.

En general, los datos analógicos son muy voluminosos y repetitivos. La mayoría de los datos analógicos consisten en una larga lista de números que se han generado. La mayoría de los registros creados por un dispositivo analógico son mediciones y la mayoría de las veces esas mediciones solo varían ligeramente de todas las demás mediciones. Por lo general, estos pequeños valores atípicos son de gran interés.
Los datos analógicos generalmente son una medida simple de algún valor físico como calor, peso, composición química o tamaño, entre otros. Cuando una medición parece estar fuera de línea, esta se interpreta como una indicación para buscar en otro lugar la causa de la variación en la medición.

Es por eso por lo que la información del metaproceso asociada con los datos analógicos a menudo es más importante que los datos analógicos en sí. Los detalles del metaproceso generalmente incluyen información como el tiempo de medición, la ubicación de la medición, la velocidad de medición, etc.

Por lo general, la información analógica se activa o se asocia con algún disparador que causa la creación de un registro analógico. La medición analógica casi siempre se realiza mecánicamente, sin ninguna intervención del usuario o procesamiento adicional.

Los puntos de datos que acompañan a los datos sin procesar capturados en el proceso de medición analógico, se denominan datos de "metaprocesos". Si bien existen diferentes tipos de modelos de metaprocesos adecuados para diferentes objetivos, este producto bruto es el más relevante para los lagos de datos. La información del metaproceso proporciona una perspectiva diferente de los datos analógicos, más que sólo mirar los datos sin procesar en sí.

DATOS DE APLICACIÓN

La segunda categoría general de datos que se encuentra en los lagos de datos son los datos de aplicación.

Los datos de aplicación se generan mediante la ejecución de una aplicación o transacción, y se envían al lago de datos. Aunque los datos de transacción son muy importantes, no son el único tipo de datos que se encuentran en el componente de aplicación del lago de datos. Los típicos datos de aplicación encontrados en el lago de datos incluyen datos de ventas, datos de pago, datos de cheques bancarios, datos de control del proceso de fabricación, datos de envío, datos de finalización del contrato, datos de gestión de inventario, datos de facturación, datos de pago de facturas, entre otros. Cuando ocurre cualquier evento relevante para el negocio, una aplicación mide el evento y se crean los datos.

La manifestación física de los datos de aplicación en el lago de datos puede tomar muchas formas, sin embargo, la forma más típica es registrar la actividad de una aplicación. Los registros pueden o no haber sido formados por una aplicación del sistema de gestión de bases de datos (DBMS). Es típico de los registros de aplicaciones tener una estructura uniforme común y repetitiva, que es más que un punto de datos analógico. El registro puede tener atributos. Uno o más de esos atributos pueden designarse como una clave. Uno o más de los atributos pueden tener un índice

independiente. Es de destacar que la estructura de los datos de la aplicación puede o no estar rigurosamente vinculada al DBMS en el que una vez se almacenaron los datos.

DATOS TEXTUALES

El tercer tipo general de datos encontrados en el lago de datos son los datos textuales.

Los datos textuales generalmente están asociados con una aplicación. Sin embargo, los datos textuales toman una forma muy diferente a los datos de aplicación. Mientras que los datos de aplicación tienen forma de registros uniformes, los datos que se encuentran en un formato de texto definitivamente no tienen forma uniforme.

Los datos textuales se denominan "datos no estructurados" porque el texto puede adoptar cualquier forma. En un texto las palabras pueden expresarse de muchas maneras diferentes. Podrían hablar en adivinanzas y parábolas. Pueden usar un idioma diferente. Su discurso puede contener jerga, vulgaridades, tener un estilo formal o incluso ser una broma interna. Naturalmente, dicho texto es extremadamente dependiente del contenido y no se busca o procesa fácilmente por medios automatizados.

Para que el texto se use analíticamente debe ser transformado. Mientras el texto esté en su forma original, solo el análisis más superficial puede hacerse contra el texto. Para que el texto sea sometido a un procesamiento analítico útil, el texto no estructurado debe pasar por un proceso conocido como desambiguación textual.

Los datos analógicos y los datos de aplicaciones rara vez tienen que pasar por un proceso similar. Debido a la uniformidad con la que se capturan los datos analógicos y los datos de la aplicación, se espera que ese tipo de puntos de datos sean analizados por una computadora. Pero si se va a realizar un análisis exhaustivo del texto, se debe pasar de su forma no estructurada de datos a través de la desambiguación textual, en cuyo punto pasa a un estado y una forma que la computadora puede analizar.

Hay dos actividades principales que se logran mediante desambiguación textual: el texto pasa de un estado no estructurado a un estado uniforme estructurado que puede ser

analizado por la computadora, y el texto tiene un contexto reconocido y asociado con el texto mismo. Si bien estas son las dos funciones principales de la desambiguación textual, existen otras funciones útiles que se logran mediante ella. La más compleja de estas actividades de desambiguación es la identificación del contexto del texto y la asociación del texto con ese contexto.

6.3. ESTANQUES DE DATOS

Para organizar los diferentes tipos de datos en una estructura que pueda analizarse, es necesario crear una estructura de datos de alto nivel dentro del lago de datos.

A medida que los datos ingresan al lago, primero llegan al estanque de datos sin procesar. El propósito del estanque de datos sin procesar es servir como celda de retención. Hay poco o ningún análisis u otra actividad organizada de los datos mientras está en el estanque de datos sin procesar. Una vez que es hora del análisis, la información en el estanque de datos sin procesar se envía a uno de los tres estanques diferentes en función del tipo de datos involucrados. Por ejemplo, los datos analógicos, de aplicación y de texto requieren un conjunto de datos único. Si bien es importante separar los tres tipos de datos, una vez dentro del estanque se lleva a cabo un procesamiento considerable. Es de destacar que dentro del estanque de datos se producen tipos muy diferentes de procesamiento o acondicionamiento de datos. Una vez finalizado el acondicionamiento, los datos en el estanque son aptos para el análisis.

Una vez que los datos han sobrevivido su vida útil en el estanque de datos, se trasladan del estanque analógico, de aplicación o textual a un estanque de datos de archivo.

DATOS DE ACONDICIONAMIENTO

A medida que los datos ingresan a los diversos estanques de origen, los datos sin procesar pasan por un proceso de acondicionamiento para preparar los datos para el procesamiento analítico. Dicho de otra manera, si los datos sin procesar no pasan por el proceso de acondicionamiento, es posible que no puedan generar valor en el análisis

empresarial. Esto se debe a que la información no está en un formato que sea fácil, o incluso posible, de estudiar. Es absolutamente obligatorio que los datos sin procesar estén acondicionados para que sean aptos para soportar el procesamiento analítico. Pero el acondicionamiento para cada tipo de estanque es muy diferente.

ESTANQUE DE DATOS SIN PROCESAR

Todo comienza en el estanque de datos sin procesar. El estanque de datos sin procesar es lo que muchas organizaciones inicialmente llaman el lago de datos. Con demasiada frecuencia, simplemente arrojan datos al lago y luego se preguntan por qué no pueden realizar ningún procesamiento analítico significativo con los datos. Para ser justos, el procesamiento analítico se puede realizar contra datos sin procesar en el lago de datos. Solo requiere un científico de datos para hacer el análisis. Pero se puede hacer un análisis de datos mucho más lúcido y eficiente con los datos una vez que han sido acondicionados.

Una vez que los datos sin procesar pasan del conjunto de datos sin formato al conjunto de datos analógicos, al conjunto de datos de aplicación o al conjunto de datos textuales, se deben eliminar los datos de origen del conjunto de datos sin formato. Los datos en bruto ya han cumplido su propósito y sería extremadamente raro que el procesamiento analítico se realice en el estanque de datos en bruto. El estanque de datos sin procesar se convierte en una "celda de retención" para una mezcla de datos.

Los datos en el estanque de datos sin procesar deben pasarse a los estanques de datos de soporte lo más rápido posible. Una medida útil de calidad para el estanque de datos en bruto es lo pequeño que es y la rapidez con que los datos salen del estanque.

ESTANQUE DE DATOS ANALÓGICOS

El estanque de datos analógicos es el lugar donde se almacenan datos analógicos. El proceso de acondicionamiento para este tipo de datos consiste principalmente en la reducción de datos: reducir el volumen de datos en el estanque analógico a un volumen de datos viable, manejable y significativo, y reestructurar los datos en el estanque.

ESTANQUE DE DATOS DE APLICACIÓN

El estanque de datos de aplicación se llena con información que proviene de la ejecución de una o más aplicaciones. Los datos de aplicación son probablemente los "más limpios" en el lago de datos porque han sido generados por una aplicación. Todos los datos en el estanque de aplicaciones están estructurados de manera uniforme y contienen valores que son relevantes para la ejecución de alguna actividad comercial. Pero los datos en el estanque de aplicaciones están notoriamente desintegrados. Si por casualidad, toda la información en este estanque proviene de una sola aplicación, los datos en este estanque en realidad pueden integrarse. Sin embargo, para las grandes corporaciones, existe una buena posibilidad de que los datos en este estanque provengan de diferentes aplicaciones. Es este origen de datos de múltiples aplicaciones lo que le dificulta el trabajo al analista.

ESTANQUE DE DATOS TEXTUALES

El estanque de datos textuales es donde se colocan los datos textuales no estructurados. El texto aquí puede venir de cualquier parte. El texto en este estanque es notoriamente difícil de analizar de manera profunda. El texto puede tener un análisis superficial sin transformación, pero para hacer un análisis profundo de los datos es necesario desambiguar el texto.
La desambiguación del texto tiene dos efectos importantes: el texto se reestructura en un formato uniforme de base de datos y el texto tiene un contexto identificado y adjunto al texto mismo.

TRANSFERENCIA DE DATOS DIRECTAMENTE EN LOS PUNTOS DE DATOS

Vale la pena señalar que los datos no tienen que pasar por el estanque de datos sin procesar, aunque casi siempre lo hacen. Si el desarrollador es sofisticado, es posible enviar los datos directamente al análogo, la aplicación o el conjunto de datos textuales. Sin embargo, la mayoría de los datos pasan por el estanque de datos sin procesar simplemente porque esa es la forma en que la mayoría de las organizaciones lo hicieron al principio.

En las etapas finales del ciclo de vida de los datos, los datos pasan del estanque de datos analógico, de aplicación o de texto al estanque de archivo.

ESTANQUE DE DATOS DE ARCHIVO

El propósito del estanque de datos de archivo es mantener datos que no se necesitan activamente para el análisis, pero que podrían ser necesarios en algún momento futuro para el análisis.

ESTRUCTURA GENÉRICA DEL ESTANQUE DE DATOS

Cada uno de los estanques de datos, a excepción del estanque de datos sin procesar, tiene algunos componentes comunes:

- Descriptor del estanque: El descriptor del estanque contiene una descripción de los contenidos externos y la manifestación del estanque, y de dónde se originaron los datos en el estanque.

- Objetivo del estanque: El objetivo del estanque es una descripción de la relación entre el negocio de la corporación y los datos dentro del estanque.

- Datos del estanque: Los datos en el estanque son simplemente los datos físicos que residen dentro del estanque.

- Metadatos del estanque: Los metadatos describen las características físicas de los datos contenidos en el estanque de datos.

- Metaproceso del estanque: La información del metaproceso es información sobre el acondicionamiento de transformación de los datos dentro del estanque de datos. Para ser útiles, los datos en el estanque deben someterse a un proceso de acondicionamiento de transformación.

- Criterios de transformación del estanque: Los criterios de transformación del estanque son documentación de cómo debe ocurrir la transformación / acondicionamiento de los datos dentro del estanque.

DESCRIPTOR DEL ESTANQUE

El descriptor del estanque tiene información acerca de:

Frecuencia de actualización o refresco: La frecuencia de actualización o refresco se refiere al ciclo con el cual los datos se envían al estanque de datos y/o la frecuencia o ciclo de refresco de datos fuera del estanque. Este puede ser un movimiento de datos programado regularmente.

Descripción de la fuente: La descripción de la fuente describe el linaje de los datos en el estanque de datos. En muchos casos, el linaje de los datos pasará por más de una fuente. Esta información de linaje es útil para el analista al determinar la idoneidad de los datos en el estanque de datos para el análisis.

Volumen de datos: El volumen de datos es una descripción general de cuántos datos hay en el estanque de datos. Los datos se miden tanto en términos de número de registros como en número de bytes. El volumen de datos influye mucho en el tipo y la profundidad del análisis que se puede hacer.

Criterio de selección. Los criterios de selección son una descripción de los criterios que se utilizaron para seleccionar los datos para su inclusión en el estanque de datos. Los criterios de selección de datos son importantes para el analista porque permiten determinar qué datos hay en el estanque y por qué están allí.

Criterios de resumen: La mayoría de las veces, los datos se resumen o se procesan a medida que pasan al estanque de datos. El resumen es una descripción de los algoritmos empleados. En algunos casos, los datos se transforman en un modelo diferente al resumen. Esta es una descripción del procesamiento algorítmico utilizado en la configuración de los datos en el estanque de datos. Los criterios de resumen son útiles para el analista para determinar cómo hacer el análisis.

Criterios de organización: Una vez que los datos se colocan en el estanque de datos, generalmente se organizan a lo largo de las líneas del objetivo del estanque. El objetivo del estanque es similar al modelo de datos del negocio. La organización de los datos puede ser rigurosa u casual, pero en cualquier caso hay una descripción de cómo se organiza exactamente el estanque. La descripción de la organización de datos es útil para el analista de negocios que intenta dar sentido al estanque de datos.

Relaciones de datos: Normalmente hay muchas relaciones de datos entre los datos encontrados en el estanque. Esta es una descripción de esas relaciones. Las relaciones de datos son útiles para el analista de negocios cuando llega el momento de hacer análisis de negocios.

OBJETIVO DEL ESTANQUE

El objetivo del estanque es el modelo básico que se utiliza para dar forma a los datos en el estanque de datos. El objetivo del estanque puede ser tan formal como un modelo de datos o puede ser tan informal como una descripción general de los datos encontrados en el estanque de datos.

El objetivo del estanque es el medio por el cual se establece una relación comercial con los datos del estanque de datos. El objetivo del estanque es invaluable para el analista de negocios en la planificación de cómo realizar un análisis. Entonces habrá, necesariamente, una relación comercial entre los elementos encontrados en el objetivo y el negocio mismo.

DATOS DE ESTANQUE

Los datos del estanque son la manifestación física de los datos en sí mismos, ya que residen en el estanque. Los datos se pueden organizar de muchas maneras según el mecanismo de almacenamiento del estanque de datos. En el mundo de Big Data, es habitual que la información se almacene en forma de "esquema de lectura". En este sistema, los datos se almacenan inicialmente en un bloque de datos. Luego, cuando se realiza una consulta con los datos, el sistema va y lee el bloque de datos y determina el esquema dentro del bloque.

Al organizar los datos de esta manera, se pueden almacenar cantidades muy grandes de datos de manera eficiente. Sin embargo, al almacenar los datos en forma de "esquema de lectura", la recuperación y el análisis de los datos pueden causar una sobrecarga significativa para el sistema. Cada vez que se accede a los datos, se debe acceder a todos los datos del estanque en una organización de datos de "esquema de lectura".

METADATOS DEL ESTANQUE

Un componente importante del estanque de datos son los metadatos que describen las características físicas de los datos que residen en el estanque.

Los metadatos dependen de los datos que existen fuera del estanque y de la organización física del estanque. Si los datos se almacenan en un DBMS estándar fuera del estanque, muchas o todas de esas características se llevarán dentro. En este caso, el analista puede esperar encontrar los mismos registros, atributos, claves e índices. Pero si los datos se almacenan en forma de documento fuera del conjunto de datos, entonces el analista puede esperar encontrar los datos organizados en un documento por organización de documentos.

Incluso en el caso de los datos almacenados en un sistema de "esquema de lectura", todavía se necesitan metadatos. Sin embargo, los datos que se organizan físicamente dentro del estanque, se describirán mediante metadatos. Sin las descripciones de metadatos, el analista tendría dificultades para descifrar cómo leer y analizar el estanque de datos.

METAPROCESO DE ESTANQUE

La descripción del metaproceso de la transformación que tiene lugar dentro del estanque de datos se encuentra en el mismo estanque. Los datos ingresan al estanque de datos en estado sin procesar. Los datos se "acondicionan" o se transforman en una forma y estructura que hace que los datos sean útiles e inteligibles para el analista.

Es de destacar que el proceso de acondicionamiento para cada estanque de datos es bastante diferente al proceso de acondicionamiento para otros estanques de datos. El estanque analógico tiene su proceso de acondicionamiento que es bastante diferente del proceso de acondicionamiento para el estanque de datos de aplicación o el estanque de datos textuales.

La información del metaproceso también puede describir el procesamiento que se ha producido fuera del estanque de datos. En ocasiones, se ha producido un procesamiento comercial significativo mucho antes de que los datos llegaran al estanque de datos. Es completamente posible que la información del metaproceso se pueda recopilar y almacenar al procesar datos. La información del metaproceso describe el proceso de acondicionamiento que es necesario para cada estanque de datos.

CRITERIOS DE TRANSFORMACIÓN DEL ESTANQUE

Los criterios de transformación son una descripción de los criterios utilizados en el proceso de transformación para el acondicionamiento de datos dentro del estanque de datos. Cada uno de los estanques de datos tiene sus propios criterios de transformación únicos. El criterio de transformación es a dónde va el analista para determinar exactamente cómo se han logrado las transformaciones.

6.4. ESTANQUE DE DATOS DE ARCHIVO

El estanque de datos de archivo se alimenta de datos del estanque de datos analógicos, el estanque de datos de la aplicación y el estanque de datos textuales.

El estanque de datos de archivo se utiliza para almacenar datos cuya vida útil probable ha disminuido. El propósito de este estanque es tener un lugar para almacenar datos que puedan tener algún uso futuro. Permitir que se eliminen datos inútiles de los estanques de datos para que el análisis en esos estanques de datos pueda proceder de manera eficiente.

CRITERIOS DE RETIRO

Existen varios criterios para la eliminación de datos de los estanques de datos analógicos, de aplicación y de texto. Algunos críticos son:

- El envejecimiento de los datos.
- La reducción de la probabilidad de uso.
- La necesidad de almacenar datos debido a la actividad litigada.
- La necesidad de almacenar datos debido a la criticidad, independientemente de la probabilidad de acceso.

ALTERACIÓN ESTRUCTURAL

A medida que se reestructuran los datos de los estanques de datos al estanque de datos de archivo, se produce un cambio estructural en los datos. Los datos en el estanque de datos de archivo tienen información de metadatos y metaprocesos directamente adjunta a los datos sin procesar. Este archivo adjunto garantiza que cuando los futuros analistas revisen los datos de archivo, no se pierda la información de metadatos y metaprocesos. Una vez que los datos se colocan en el estanque de datos de archivo, a menudo es útil indexar los datos de forma independiente para que los futuros analistas puedan encontrar datos de manera eficiente.

6.5. CARACTERÍSTICAS BÁSICAS ESPERADAS EN UN SISTEMA DE METADATOS

Existen seis (6) funcionalidades principales que deberían ser proporcionadas idealmente por el sistema de metadatos de un lago de datos [4].

1. El enriquecimiento semántico (SE): También llamado perfil semántico, consiste en generar una descripción del contexto de los datos, por ejemplo, con etiquetas, para hacerlos más interpretables y comprensibles. Se realiza utilizando bases de conocimiento como ontologías. La anotación semántica juega un papel clave en la

explotación de datos, al resumir los conjuntos de datos contenidos en el lago para que sean más comprensibles para el usuario. También se puede utilizar como base para identificar enlaces de datos. Por ejemplo, los datos asociados con las mismas etiquetas pueden considerarse vinculados.

2. La indexación de datos (DI): consiste en establecer una estructura de datos para recuperar conjuntos de datos basados en características específicas (palabras clave o patrones). Esto requiere la construcción de índices hacia adelante o invertidos. La indexación permite optimizar las consultas de datos en el lago mediante el filtrado de palabras clave. Es particularmente útil para la gestión de datos textuales, pero también se puede utilizar en un contexto de datos semiestructurado o estructurado.

3. La generación y conservación de enlaces (LG): es el proceso de detectar relaciones de similitud o integrar enlaces preexistentes entre conjuntos de datos. La integración de enlaces de datos se puede utilizar para ampliar el rango de posibles análisis del lago recomendando datos relacionados con los de interés para el usuario. Los enlaces de datos también se pueden usar para identificar grupos de datos, es decir, grupos donde los datos están fuertemente vinculados entre sí y son significativamente diferentes de otros datos.

4. El polimorfismo de datos (DP): se define como el almacenamiento de múltiples representaciones de los mismos datos. Cada representación corresponde a los datos iniciales, modificados o reformateados para una necesidad específica. Por ejemplo, un documento de texto puede representarse sin palabras vacías o como una bolsa de palabras. En el contexto de los lagos de datos, es esencial estructurar al menos parcialmente los datos no estructurados para permitir su análisis automatizado. Almacenar simultáneamente varias representaciones de los mismos datos evitan notablemente los preprocesos repetidos y, por lo tanto, aceleran los análisis.

5. El control de versiones de datos (DV): se refiere a la capacidad del sistema de metadatos para admitir cambios de datos mientras se conservan los estados anteriores. Esta capacidad es esencial en los lagos de datos, ya que garantiza la reproducibilidad de los análisis y apoya la detección y corrección de posibles errores

o inconsistencias. El control de versiones también permite soportar una evolución ramificada de datos, especialmente en su esquema.

6. El seguimiento de uso (UT): registra las interacciones entre los usuarios y el lago de datos. Las interacciones son generalmente operaciones de creación, actualización y acceso de datos.

La integración de esta información en el sistema de metadatos permite comprender y explicar posibles inconsistencias en los datos. También se puede usar para administrar datos confidenciales, detectando intrusiones.

El seguimiento del uso y el control de versiones de los datos están estrechamente vinculados, porque las interacciones conducen en algunos casos a la creación de nuevas versiones o representaciones de los datos. Por lo tanto, tales características a menudo se integran juntas en un módulo de seguimiento de procedencia. Sin embargo, todavía consideramos que siguen siendo características diferentes ya que no se proponen sistemáticamente juntos.

A continuación se realiza una comparación entre sistemas de metadatos, revisando dos (2) de sus tipos: modelos de metadatos e implementaciones de lago de datos.

Los modelos de metadatos se refieren a sistemas conceptuales para organizar metadatos. Tienen la ventaja de ser más detallados y más fácilmente reproducibles que las implementaciones del lago de datos, que se encuentran en un nivel más operativo [5].

Las implementaciones del lago de datos se centran en la operación y la funcionalidad, con pocos detalles sobre la organización conceptual de los metadatos.

SISTEMA	TIPO	SE	DI	LG	DP	DV	UT
SPAR (Fauduet and Peyrard, 2010)	DL	X	X	X			X
Alrehamy and Walker (2015)	DL	X		X			
Terrizzano et al. (2015)	DL	X	X			X	X
Constance (Hai et al., 2016	DL	X	X				
GEMMS (Quix et al., 2016)	MM	X					
CLAMS (Farid et al., 2016)	DL	X					
Suriarachchi and Plale (2016)	MM				X		X
Singh et al. (2016)	DL	X	X	X	X		
Farrugia et al. (2016)	DL			X			
GOODS (Halevy et al., 2016)	DL	X	X	X		X	X
CoreDB (Beheshti et al., 2017)	DL		X				X
Ground (Hellerstein et al., 2017)	MM	X	X			X	X
KAYAK (Maccioni and Torlone, 2018)	DL	X	X	X			
CoreKG (Beheshti et al., 2018)	DL	X	X	X	X		X
Diamantini et al. (2018)	MM	X		X	X		
MEDAL (Sawadogo et al., 2019)	MM	X	X	X	X	X	X

DL: Implementación de Data Lake
MM: Modelo de Metadatos

6.6. ARQUITECTURAS ALTERNAS

Rajesh Nadipalli [6] presenta la arquitectura del lago de datos de Amazon Web Services (AWS) con cuatro zonas: ingestión, almacenamiento, procesamiento y gobierno y seguridad:

- Zona de ingestión: Allí se cargan los datos sin procesar.
- Zona de almacenamiento: Aquí se almacenan los datos brutos ingeridos.
- Zona de procesamiento: Aquí se procesan los datos cuando es requerido.
- Zona de gobierno y seguridad: Permite controlar la seguridad de los datos, la calidad de los datos, la gestión de metadatos y el ciclo de vida de los datos.

Pradeep Menon [7] separa la zona de procesamiento de datos en zonas de procesamiento por lotes y en tiempo real:

- Motor de procesamiento por lotes: Procesa los datos sin procesar en algo que los usuarios pueden consumir, es decir, una estructura que se pueda usar para informar al usuario final. ES llamado almacén de datos procesados.

- Motor de procesamiento en tiempo real: Toma la transmisión de datos y también los procesa. Todos los datos en esta arquitectura están catalogados y curados.

Alice LaPlante y Ben Charma [8] proponen separar las zonas de procesamiento y almacenamiento en una zona de datos refinada, una zona de datos de confianza y una zona de pruebas de descubrimiento:

- Zona refinada: Permite integrar y estructurar datos.

- Zona de datos de confianza: Almacena todos los datos limpios.

- Zona de descubrimiento: Allí se mueven los datos para el análisis exploratorio.

Franck Ravat y Yan Zhao [9] [10] proponen una arquitectura funcional que contiene cuatro (4) zonas esenciales, y cada zona, excepto la zona de gobierno, tiene un área de tratamiento y un área de almacenamiento de datos que almacena el resultado de los procesos:

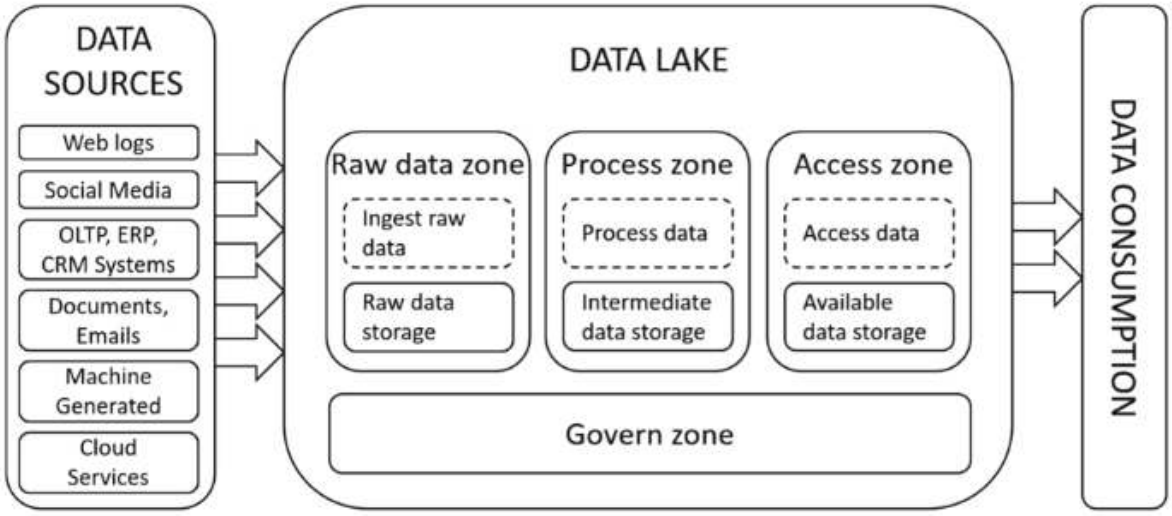
- Zona de datos sin procesar: todos los tipos de datos se ingieren sin procesamiento y se almacenan en su formato nativo. La ingestión puede ser por lotes, en tiempo real o híbrida. Esta zona permite a los usuarios encontrar la versión original de los datos para sus análisis para facilitar los tratamientos posteriores. El formato de datos sin procesar almacenado puede ser diferente del formato de origen.

- Zona de proceso: en esta zona, los usuarios pueden transformar datos de acuerdo con sus requisitos y almacenar todos los datos intermedios. El procesamiento de datos incluye procesamiento por lotes y/o en tiempo real. Esta zona permite a los

usuarios procesar datos (selección, proyección, unión, agregación, etc.) para su análisis de datos.

- Zona de acceso: la zona de acceso almacena todos los datos disponibles para el análisis de datos y proporciona el acceso a los datos. Esta zona permite el consumo de datos de autoservicio para diferentes análisis (informes, análisis estadísticos, análisis de inteligencia empresarial, algoritmos de aprendizaje automático).

- Zona de gobernanza: la gobernanza de datos se aplica en todas las demás zonas. Está a cargo de garantizar la seguridad de los datos, la calidad de los datos, el ciclo de vida de los datos, el acceso a los datos y la gestión de metadatos.



7. HERRAMIENTAS UTILIZADAS EN EL ENTORNO DE LOS DATA LAKES

Hay una variedad de herramientas que admite el entorno del lago de datos. Cada uno proporciona una funcionalidad diferente que se necesita en este. Se mencionarán las más destacadas.

VISUALIZACIÓN

La visualización es la tecnología que toma los datos, los organiza. Al convertir los detalles de una base de datos en una visualización, la organización puede ver de inmediato patrones y tendencias que de otro modo no serían obvios.

La visualización es especialmente útil para la gestión no técnica. En muchos casos, los ejecutivos de una organización no pueden entender lo que se dice a menos que se visualicen los datos. La tecnología de visualización puede organizar los datos en una variedad de formas.

Para que sea eficaz, los datos que ingresan a una visualización deben organizarse primero en un formato de base de datos. La mayoría de la tecnología de visualización requiere que los datos en los que opera se almacenen en un formato de base de datos relacional.

BUSCAR Y CALIFICAR

Otra tecnología útil y sofisticada es la tecnología de búsqueda y calificación. Algunas tecnologías de búsqueda son bastante simples, mientras que otras son muy sofisticadas. La tecnología de búsqueda y calificación puede hacer búsquedas sofisticadas donde los datos pueden estar organizados de manera menos óptima, como en el caso de los datos textuales. Una de las formas sofisticadas de la tecnología de búsqueda es el aprendizaje automático y la tecnología de búsqueda de conceptos. En la tecnología de aprendizaje automático y búsqueda de conceptos, los documentos de texto se pueden leer y calificar.

DESAMBIGUACIÓN TEXTUAL

La tecnología más útil en el estanque de datos textuales es la tecnología conocida como desambiguación textual. En la tecnología de desambiguación textual, la narración textual en bruto se lee y se convierte a un formato de base de datos estándar. Además, el contexto del texto se identifica y se escribe junto con el texto.

La desambiguación textual es una tecnología compleja. Se trata de lenguaje y el lenguaje es intrínsecamente complicado. Para aquellas organizaciones que realizan análisis textuales serios, la desambiguación textual es una necesidad absoluta.

ANÁLISIS ESTADÍSTICO

El análisis estadístico es otra tecnología que es bastante útil para leer grandes cantidades de datos y hacer análisis estadísticos sofisticados de los datos. El análisis estadístico implica no solo el cálculo de números analíticos, sino la visualización gráfica de esos números de una manera significativa.

PROCESAMIENTO ETL CLÁSICO

El ETL clásico es útil para leer e integrar datos de aplicaciones y, por lo tanto, el proceso de transformación. El procesamiento ETL clásico lee datos basados en aplicaciones y los convierte en datos corporativos que se han integrado.

8. APLICACIONES DE LOS DATA LAKES

Las aplicaciones que se le han dado a los lagos de datos son principalmente de tipo empresarial, y tienen sentido en la medida en que brinden valor empresarial a las organizaciones. Los diferentes estanques de datos tienen potencial para proporcionar valor empresarial, pero el valor proporcionado por cada conjunto de datos y la forma en que se proporciona dicho valor son muy diferentes. Se analizará el valor que se genera desde cada estanque del lago de datos.

VALOR EMPRESARIAL EN EL ESTANQUE DE DATOS ANALÓGICOS

El estanque de datos analógicos puede proporcionar valor empresarial de una de dos maneras. Puede haber un puñado de registros encontrados o puede haber patrones de datos que se desarrollan en una vista de muchos registros de datos.

Considere una empresa que fabrica airbags para automóviles. Si un airbag no funciona correctamente, puede haber consecuencias muy graves. Suponga que ocurre un accidente donde no se dispara una bolsa de aire. El investigador del accidente encuentra al fabricante del airbag. Luego, el investigador determina que el airbag fue fabricado en marzo de 1995 en las instalaciones de Phoenix, Arizona. La compañía ahora revisa sus datos analógicos y encuentra todas las otras bolsas de aire que se fabricaron en marzo y abril de 1995 y alerta a los propietarios de los automóviles que tienen estas bolsas de aire para que revisen sus bolsas de aire, evitando así una consecuencia potencialmente grave. En este caso, se examinaron los datos analógicos para encontrar un puñado de registros que tenían consecuencias potencialmente muy graves.

Otro valor empresarial de los datos analógicos es mirar rápidamente a través de grandes vistas de datos. Un día, la gerencia desea repensar la forma en que un airbag es fabricado porque hay una nueva tecnología que activa un airbag de manera más eficiente y segura. El fabricante observa una gran cantidad de datos análogos para determinar cuántas bolsas de aire hay con el mecanismo de disparo más antiguo. Como otro ejemplo de encontrar algunos registros valiosos, considere los registros detallados del registro

de llamadas telefónicas. Un día el gobierno encuentra llamadas telefónicas entre terroristas. Puede haber millones y millones de registros detallados de llamadas telefónicas, pero solo unos pocos son de terroristas. No hay duda sobre el valor de poder identificar a los terroristas y así prevenir los actos de terrorismo. En este caso, se examinan muchos, muchos registros con la esperanza de encontrar solo un puñado de registros.

Mirar a través de vistas de datos es un asunto completamente diferente. En lugar de buscar algunos puntos de muchos, el analista busca patrones de datos que se manifiestan en muchos registros. Los patrones de datos se detectan no solo mediante la búsqueda de registros, sino también mediante el uso de la información de su metaproceso junto con los registros mismos.

VALOR EMPRESARIAL EN EL ESTANQUE DE DATOS DE APLICACIÓN

Encontrar valor empresarial del estanque de datos de aplicación es una propuesta diferente. Algunos ejemplos típicos de encontrar valor empresarial son localizar un recibo particular o la determinación del costo promedio de los envíos para 1999. Suponga que la organización está pasando por una auditoría y está buscando documentación de un año anterior. El documento es necesario para demostrar a un auditor un elemento de gastos. Los sistemas operativos solo se remontan tres años atrás, pero la auditoría es de hace cinco años. La organización busca el estanque de datos de su aplicación para encontrar el recibo. En este caso, hubo una búsqueda en muchos documentos con la esperanza de encontrar solo uno. En otra circunstancia, la gerencia piensa que los costos de envío están aumentando demasiado rápido. Para obtener una perspectiva histórica de los costos, la administración se remonta a 1999 para calcular los costos de envío. Encuentran esos costos de envío en el estanque de datos de la aplicación. Para determinar los costos de envío anuales, se debe hacer un cálculo utilizando muchos documentos.

VALOR EMPRESARIAL EN EL ESTANQUE DE DATOS TEXTUALES

Sin embargo, un tercer tipo de valor empresarial puede derivarse del estanque de datos textuales. Supongamos que se ha acordado un precio para un pedido. Sin embargo, la

única documentación está por escrito, en una carta en papel. La organización busca en todo el estanque de datos textuales para encontrar un documento.

Otro tipo de valor empresarial que puede derivarse del estanque de datos textuales es determinar el sentimiento del cliente. El sentimiento del cliente se expresa de muchas maneras: a través de tweets, correos electrónicos y otras formas de narración.

La organización lee y almacena estos documentos en su estanque de datos textuales, que luego pasa estos documentos a través de la desambiguación textual y crea una base de datos que puede analizarse, lo que facilita determinar el sentimiento del cliente. El sentimiento del cliente se mide mirando muchos documentos, leyendo y desambiguando el contenido de los documentos, y colocando los resultados en una base de datos, donde se puede realizar el análisis. Conocer el sentimiento del cliente es algo extremadamente valioso para el negocio.

PORCENTAJE DE REGISTROS QUE TIENEN VALOR EMPRESARIAL

Otra forma interesante de ver el valor empresarial proporcionado por los diferentes estanques de datos es a través del porcentaje de registros que tienen valor empresarial. Algunos puntos de datos tienen registros que tienen un porcentaje muy alto de dicho valor, mientras otros no tanto.

Como un ejemplo tenemos las llamadas telefónicas. En los Estados Unidos cada día, se realizan millones de llamadas telefónicas. Si una persona buscaba llamadas telefónicas hechas por terroristas, es seguro decir que solo hay un puñado de puntos relevantes. De hecho, en un día cualquiera puede que no haya llamadas telefónicas hechas por terroristas. Cuando observa el porcentaje de llamadas telefónicas terroristas realizadas cada día frente al número total de llamadas, el porcentaje es muy bajo. Quizás el porcentaje sea tan bajo como .0000001%. Y los mismos porcentajes muy bajos de registros que tienen valor empresarial son válidos para cosas como cintas de registro, registros de flujo de clics y muchos otros datos.

Revisando otro ejemplo como los datos textuales, se ve que estos se recopilan de lugares como conversaciones del centro de llamadas, comentarios de los clientes, entre otros.

Cada llamada telefónica representa las inquietudes o mensajes de un cliente. El contenido de cada llamada telefónica tiene un valor empresarial real. Para la mayoría de los datos textuales, el 100% de los datos tiene valor empresarial. Es cierto que algunas conversaciones telefónicas tienen más valor que otras. Pero cada conversación telefónica tiene algún valor empresarial.

9. CONCLUSIONES

- ✓ Los lagos de datos o Data Lakes son una evolución de los almacenes de datos o Data Warehouses, que pretenden ampliar el abanico de posibilidades en el análisis de datos de cualquier tipo con el fin de generar valor a las organizaciones que los producen.

- ✓ Los lagos de datos brindan organización y la posibilidad de gobernanza de una manera eficiente a la Big Data.

- ✓ Los lagos de datos son una tecnología que aún está en proceso de construcción, por lo que aunque existe una arquitectura inicialmente planteada, esta ha servido de base para que nuevas arquitecturas ajusten, en mayor o menor alcance, los elementos base para su funcionamiento, y puedan así solucionar problemas planteados desde diversas perspectivas.

- ✓ En todas las arquitecturas diferentes que se plantean, hay unos elementos comunes que son fundamentales para el funcionamiento del lago de datos. Estos son los metadatos. Aunque tengan diferentes configuraciones internas, los metadatos se constituyen en la principal herramienta de filtrado y organización a la hora de realizar análisis o transformaciones en los datos que residen en el lago.

- ✓ Las diferentes arquitecturas descritas comparten el criterio de estanques especializados, ya sea para una categoría de datos o una funcionalidad. Esto indica que, aunque en principio todos los datos de cualquier tipo caen en esta laguna, cada categoría tiene su forma particular de ser procesada y transformada en información de valor para la organización.

- ✓ Al hablar de herramientas utilizadas en el entorno de los lagos de datos, se hace referencia a las metodologías que pueden ser utilizadas para cada uno de los procesos que se llevan a cabo allí, tales como la carga de datos, la clasificación de estos, la transformación requerida para su tratamiento, el análisis, y hasta la manera

de mostrar los resultados a los interesados. De estas herramientas hay múltiples opciones que se utilizarán dependiendo del contexto de los datos y los requisitos de los interesados del proceso.

- ✓ En lo que respecta a las aplicaciones de los lagos de datos, se tiene que su aplicación es básicamente de tipo empresarial, pero allí, dependiendo de los requisitos de los interesados, su aplicación radica en el valor real que aporte a ellos, ya sea en la optimización de los procesos de las organizaciones, en las áreas de investigación, innovación y desarrollo, o en el área comercial específicamente, y se deja claro que dicho valor dependerá principalmente del tipo de datos que se alberguen en el lago de datos.

BIBLIOGRAFÍA

[1] H. Fang, «Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem,» de *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, Shenyang, 2015.

[2] P. P. Khine y Z. S. Wang, «Data lake: a new ideology in big data era,» de *ITM web of conferences*, 2018.

[3] B. Inmon, Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump, 1 ed., Basking Ridge: Technics Publications, 2016.

[4] P. N. Sawadogo, E. Scholly, C. Favre, E. Ferey, S. Loudcher y J. Darmont, «Metadata systems for data lakes: models and features,» de *New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science*, vol. 1064, Cham, Springer, 2019, pp. 440-451.

[5] P. Sawadogo, T. Kibata y J. Darmont, «Metadata management for textual documents in data lakes,» de *21st International Conference on Enterprise Information Systems (ICEIS 2019)*, Heraklion, 2019.

[6] R. Nadipalli, Effective Business Intelligence with QuickSight, Birmingham: Packt Publishing Ltd, 2017.

[7] P. Menon, «Medium.com,» 5 Julio 2017. [En línea]. Available: https://medium.com/@rpradeepmenon/demystifying-data-lake-architecture-30cf4ac8aa07.

[8] A. LaPlante y B. Sharma, Architecting data lakes: data management architectures for advanced business use cases, Sebastopol: O'Reilly Media, 2016.

[9] F. Ravat y Y. Zhao, «Data lakes: Trends and perspectives,» de *Database and Expert Systems Applications. DEXA 2019. Lecture Notes in Computer Science*, vol. 11706, Cham, Springer, 2019, pp. 304-313.

[10] F. Ravat y Y. Zhao, «Metadata management for data lakes,» de *New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science*, vol. 1064, Cham, Springer, 2019, pp. 37-44.

[11] C. Walker y H. Alrehamy, «Personal Data Lake with Data Gravity Pull,» de *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, Dalian, 2015.

[12] R. Hai, S. Geisler y C. Quix, «Constance: An intelligent data lake system,» de *Proceedings of the 2016 International Conference on Management of Data*, New York, 2016.

[13] M. Jarke y C. Quix, «On warehouses, lakes, and spaces: the changing role of conceptual modeling for data integration,» de *Conceptual Modeling Perspectives*, Cham, Springer, 2017, pp. 231-245.

[14] A. Beheshti, B. Benatallah, Q. Z. Sheng y F. Schiliro, «Intelligent Knowledge Lakes: The Age of Artificial Intelligence and Big Data,» de *Web Information Systems Engineering. WISE 2020. Communications in Computer and Information Science*, vol. 1155, Singapore, Springer, 2020, pp. 24-34.

[15] J. Kachaoui y A. Belangour, «From Single Architectural Design to a Reference Conceptual Meta-Model: An Intelligent Data Lake for New Data Insights,» *International Journal of Emerging Trends in Engineering Research,* vol. 8, nº 4, pp. 1460 - 1465, 2020.

[16] H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes y J. Riekki, «Implementing big data lake for heterogeneous data sources,» de *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, Macao, 2019.

[17] C. Giebler, C. Gröger, E. Hoos, H. Schwarz y B. Mitschang, «Leveraging the Data Lake: Current State and Challenges,» de *Big Data Analytics and Knowledge Discovery. DaWaK 2019. Lecture Notes in Computer Science*, Cham, 2019.

[18] V. Theodorou, R. Hai y C. Quix, «A Metadata Framework for Data Lagoons,» de *European Conference on Advances in Databases and Information Systems*, Cham, 2019.