

COMPARATIVA DE RAPIDEZ Y SIMPLICIDAD DE USO DE UN DATAWAREHOUSE Y DATALAKE

Balaguer Valles, Angela Lessly (2016054494), Huallpa Castro, Leydi
Katherine (2015053230), Pilco Quispe, Mireya Flavia (2015053234),
Salamanca Contreras, Fiorella Rosmery (2015053237)

Tacna, Perú

Abstract

The discussion Data Lake vs. Data Warehouse is something very common among those companies that are preparing to implement big data solutions. Quickly the conversation about data and analysis in the big data field takes us to the Data Lake or data lake, but very often the companies do not quite understand what this means and what are the differences between Data Lake vs. Data Warehouse.

1. Resumen

La discusión Data Lake vs Data Warehouse es algo muy común entre aquellas empresas que se disponen a implantar soluciones de big data. Rápidamente la conversación sobre datos y análisis en el ámbito de big data nos lleva al Data Lake o lago de datos, pero muy a menudo las empresas no acaban de entender bien qué es lo que esto significa y cuáles son las diferencias entre Data Lake vs Data Warehouse.

2. Introduccion

Actualmente trabajar con cantidades enormes de datos empieza a ser la norma más que la excepción y, es cada vez más necesario buscar una solución más eficiente para almacenar y procesar grandes volúmenes de información.

El enfoque tradicional del DataWarehouse/Business Intelligence ha hecho un gran trabajo para simplificar el acceso a los datos y la presentación de informes, permitiendo combinar datos de muchas fuentes, con el fin de responder a las preguntas que una organización puede tener.

Los datos son la clave para entender los patrones de tus clientes, competidores y mercados. Sólo mediante el análisis de esta información se pueden tomar decisiones y llevar a cabo las acciones adecuadas.

Por ello, el reto para muchas de las compañías actuales es Integrar, Gestionar y Distribuir sus datos a aquellos que los necesitan en el menor tiempo posible, apareciendo en los últimos años el concepto de Data Lake.

3. Objetivo

- Objetivo 1:
- Objetivo 2:
- Objetivo 3:

4. Marco Teorico

4.1. DATA WAREHOUSE

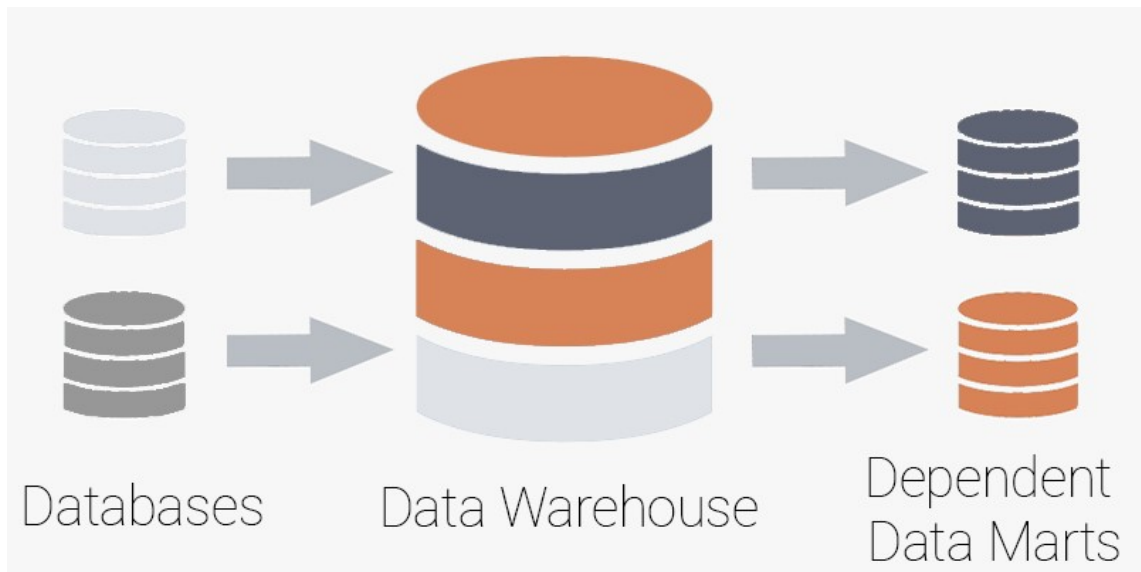
Los profesionales en este ámbito seguro que conocen este término y, cuando se empieza a hablar de soluciones de Big Data con los clientes, la conversación, habitualmente, se convierte en una discusión sobre Data Lakes. Sin embargo, a menudo encuentro que los clientes o no han escuchado el término o realmente no entienden bien lo que significa. De hecho, creo que hay cierta confusión en ocasiones entre Data Warehouse y Data Lake, por eso en este post quería enumerar brevemente las principales diferencias.

Pero inicialmente creo interesante definir estos términos, para verlas más fácilmente.

Data Warehouse es el repositorio central de los datos de una empresa provenientes de diferentes fuentes. Se guardan los datos actuales y su histórico, y se utilizan para la creación de informes y análisis de tendencias. Algunas de sus características son:

- Representa una foto abstracta de la organización del negocio por diferentes áreas.

- Sus datos están muy estructurados y organizados.
- No tiene datos cuyo uso no haya sido definido previamente.



4.2. PRINCIPALES VENTAJAS DE UN DATA WAREHOUSE

Estas son las principales ventajas que se pueden encontrar en la implantación de un Data Warehouse en el proceso de gestión del dato en tu negocio:

- Facilita la toma de decisiones basadas en datos, en cualquier área funcional de la empresa, ya que te proporciona información integrada y global del negocio.
- La información se convierte en un valor añadido para cualquier negocio, gracias a que permite aplicar técnicas estadísticas de análisis y modelización que ayudan a encontrar relaciones ocultas entre los datos almacenados.
- Te permite de manera sencilla aprender de los datos del pasado y predecir situaciones futuras para diferentes escenarios.
- Simplifica la implantación de sistemas de gestión integral de la relación con el cliente, dentro de la empresa.



- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares.
- Es un sistema especialmente útil para el medio y el largo plazo.
- Aumenta la productividad de las empresas de manera muy sustancial.
- Te permite realizar planes de una manera mucho más efectiva.
- Permite la integración de todas las herramientas corporativas. Por ejemplo, nosotros en Artyco integramos toda la información que recogemos a través de todas nuestras aplicaciones (monitorización web, crm, wifi tracking, campañas...) en un Data Warehouse, de donde sacar la información necesaria ante consultas determinadas.
- Para trabajar de manera correcta un Data Warehouse, es preciso que todos los componentes de la organización hablen el mismo lenguaje, es decir, que todos llamen a las cosas por su nombre. De este modo, gracias al Data Warehouse se pueden unificar conceptos.

4.3. *DESVENTAJAS DE UN DATA WAREHOUSE*

- No es muy útil para la toma de decisiones en tiempo real debido al largo tiempo de procesamiento que puede requerir. En cualquier caso la tendencia de los productos actuales (junto con los avances del hardware) es la de solventar este problema convirtiendo la desventaja en una ventaja.
- Requiere de continua limpieza, transformación e integración de datos.
- Mantenimiento.
- En un proceso de implantación puede encontrarse dificultades ante los diferentes objetivos que pretende una organización.
- Una vez implementado puede ser complicado añadir nuevas fuentes de datos.
- Requieren una revisión del modelo de datos, objetos, transacciones y además del almacenamiento.
- Tienen un diseño complejo y multidisciplinar.
- Requieren una reestructuración de los sistemas operacionales.
- Tienen un alto coste.
- Requieren sistemas, aplicaciones y almacenamiento específico.

Las empresas que utilizan data warehouse son fundamentalmente aquellas que manejan grandes volúmenes de datos relativos a clientes, compras, marketing, transacciones, operaciones. como lo son las empresas de telecomunicaciones, transporte, Turismo, fabricación de bienes de consumo masivo etc.

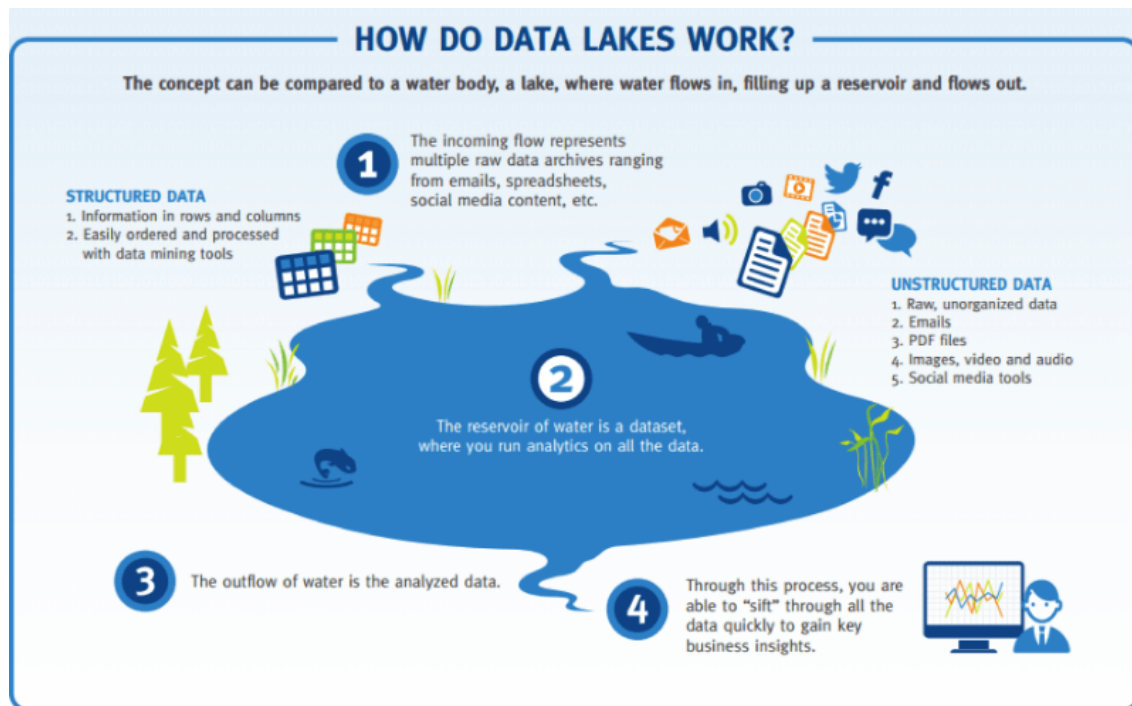
4.4. *DATA LAKE*

Data Lake es un término cuya primera definición o uso se atribuye a James Dixon que decía, “if you think of a datamart as akin to a bottle of water. . .”cleansed, packaged and structured for easy consumption” the data lake is more like a body of water in its natural state. Data flows from the streams (the source systems) to the lake. Users have access to the lake to examine, take samples or dive in.”

Sus principales funciones son la gestión de la ingesta de datos, su almacenamiento y procesamiento posterior y, por último, el acceso a los mismos.

Algunas de sus características son:


- Contiene todos los datos de las fuentes originales, sin rechazar ningún tipo de dato.
- Los datos se almacenan sin transformar o apenas transformados.
- Los datos se transforman y se aplica un esquema sólo para satisfacer las necesidades de análisis.



Es importante saber que al igual que cuando hablamos de Data Warehouse, por detrás hay una solución que soporta el modelo (Teradata, Oracle Exadata, SAP Hana, Microsoft SQL Server...) y muy habitualmente detrás de un Data Lake lo que está es la infraestructura del sistema de archivos

HDFS (Hadoop Distributed File System) que utiliza Hadoop, y cuando hablamos de Hadoop en entornos corporativos generalmente hablamos de alguna de sus soluciones comerciales tales como Cloudera, Hortonworks, MapR, IBM o Pivotal, las 5 opciones más destacadas actualmente.

Tras las definiciones anteriores creo que es fácil resaltar alguna de las principales diferencias entre ambos conceptos

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals		ta scientists et. al.

En definitiva es importante saber que aun siendo ambos conceptos, Data Warehouse y Data Lake, repositorios de información, un Data Lake no es una nueva versión 2.0 de un Data Warehouse ni su remplazo.

De hecho, se pueden complementar muy bien, diseñando una arquitectura de datos moderna, que permita seguir a las organizaciones aprovechando sus inversiones en su Data Warehouse, mientras que empiezan a recoger en su Data Lake, todos los datos que han sido ignorados o desechados anteriormente.

4.5. PRINCIPALES BENEFICIOS DE UN DATA LAKE

Un Data Lake tiene muchas ventajas. Las más destacables son estas:

- El Data Lake permite centralizar todos los datos en un mismo lugar, vengan de la fuente que vengan. Una vez incluidas en su silo correspon-

diente de información, pueden ser procesadas a través de herramientas de Big Data. Muchas veces, en esa disparidad de información, habrá datos que requieran un tratamiento especial en cuanto a seguridad. Gracias al Data Lake, este aspecto se puede solventar.

- Puede que la fuente original del dato esté obsoleta o se haya desactivado, sin embargo, su contenido puede que siga siendo valioso para el análisis. A través del Data Lake, puedes acceder a dicha información.
- Todo dato que llegue al Data Lake puede ser normalizado y enriquecido.
- Los datos se preparan en función de la necesidad del momento. Esto permite reducir considerablemente los costes y los tiempos. En el Data Warehouse, por ejemplo, es necesaria dicha preparación.
- Se puede acceder a la información y enriquecerla desde cualquier punto del planeta, por cualquier usuario autorizado por el Data Lake. Esto ayuda a la organización a recopilar más fácilmente los datos necesarios para la toma de decisiones.
- Un Data Lake pone la información en manos de un mayor número de personas dentro de cualquier organización, aprovechándose mejor la empresa de ese conocimiento que adquieren dichos individuos.

4.6. Características de un data lake

Para ser clasificado como un data lake, un repositorio de datos grandes debe exhibir tres características clave:

- Un único repositorio compartido de datos, normalmente almacenado en el Sistema de archivos distribuido (DFS). Los lagos de datos de Hadoop conservan los datos en su forma original y capturan los cambios a los datos y la semántica contextual a lo largo del ciclo de vida de los datos. Este enfoque es especialmente útil para las actividades de cumplimiento y auditoría interna. Esta es una mejora con respecto al EDW tradicional, donde si los datos han sufrido transformaciones, agregaciones y actualizaciones, es difícil juntar datos cuando es necesario, y las organizaciones tienen dificultades para determinar la procedencia de los datos.



- Incluye funcionalidades de orquestación y programación de trabajos (por ejemplo, a través de YARN). La ejecución de la carga de trabajo es un requisito previo para Hadoop empresarial. YARN proporciona administración de recursos y una plataforma central para entregar herramientas consistentes de operaciones, seguridad y control de datos en los clústeres de Hadoop, asegurando que los flujos de trabajo analíticos tengan acceso a los datos y la potencia informática que requieren.
- Contiene un conjunto de aplicaciones o flujos de trabajo para consumir, procesar o actuar sobre los datos. El fácil acceso de los usuarios es una de las características de un data lake, debido a que las organizaciones conservan los datos en su forma original. Ya sea estructurado, no estructurado o semiestructurado, los datos se cargan y almacenan tal cual. Los propietarios de datos pueden entonces consolidar datos de clientes, proveedores y operaciones, eliminando barreras técnicas e incluso políticas para compartir datos.

Los data lake son cada vez más importantes para las estrategias de datos empresariales. Los datos de los lagos responden mejor a las realidades de los datos actuales: volúmenes y variedades de datos mucho mayores, mayores expectativas de los usuarios y la rápida globalización de las economías.

4.7. DATA LAKE VS DATA WAREHOUSE

Las divergencias entre Data Lake vs Data Warehouse pueden entenderse mejor repasando algunos de los puntos diferenciadores clave de un lago de datos y el modo en que contrastan con el enfoque del almacén de datos. Se trata de los siguientes:

- El lago de datos conserva todos los datos, a diferencia del almacén de datos, donde se dedica una parte importante de tiempo a decidir qué datos incluir y no incluir en el almacén
- Un Data Lake admite todos los tipos de datos, independientemente de su tipo, formato o procedencia y sin necesidad de normalizar su estructura. La información se mantiene en su forma original y solo se transforma cuando se va a consumir.
- El Data Lake puede nutrir a todos los usuarios de la organización, incluyendo a esos perfiles técnicos con exigencias de análisis más avanzadas, que son quienes recurren a capacidades como análisis estadístico y modelado predictivo.
- A diferencia del Data Warehouse, el Data Lakes se adapta fácilmente a los cambios. El diseño del almacén es un proceso complejo y, la actualidad de los negocios, en ocasiones no puede esperar tanto tiempo. Para esas circunstancias, asegura la adaptabilidad necesaria para entregar respuestas más rápidas.

Debido a que los lagos de datos contienen todos los datos y tipos de datos, y dado que permite a los usuarios acceder a los datos antes de que se hayan transformado, depurado y estructurado, también hace posible que se obtengan resultados más rápido de lo que sería posible con un enfoque tradicional de almacenamiento de datos.

Tanto los data lakes como los almacenes de datos se utilizan de forma generalizada para almacenar big data, pero no son términos intercambiables. Un data lake es un enorme conjunto de datos en bruto cuya finalidad no se ha definido todavía. Un almacén de datos es un repositorio de datos filtrados y estructurados que ya han sido procesados para una finalidad concreta.

La gente suele confundir estos dos tipos de almacenamiento de datos, cuando en realidad son mayores sus diferencias que sus semejanzas. A decir verdad, la única similaridad real entre ambos es su máxima finalidad, que es almacenar datos.

La diferencia es importante, porque están pensadas para objetivos distintos y exigen perspectivas diferentes para optimizarlas correctamente. Mientras que a una empresa le convendrá más tener un data lake, para otra resultará más oportuno disponer de un almacén de datos.

- Cuatro diferencias principales entre un data lake y un almacén de datos

Existen varias diferencias entre un data lake y un almacén de datos. Los principales diferenciadores son la estructura de los datos, los usuarios ideales, los métodos de procesamiento y la finalidad general de los datos.

	Data lake	Almacén de datos
Estructura de datos	En bruto	Procesados
Finalidad de los datos	Por determinar	Actualmente en uso
Usuarios	Científicos de datos	Profesionales corporativos
Accesibilidad	Muy accesible y rápido de actualizar	Más complicado y caro de realizar cambios

- Estructura de datos: En bruto frente a procesados

Los datos en bruto son datos que no aún no han sido procesados para ninguna finalidad. Quizá la principal diferencia entre los data lakes y los almacenes de datos sea la diversa estructura existente entre los datos en bruto y los procesados. En líneas generales un data lake almacena datos en bruto, sin procesar, mientras que un almacén guarda datos procesados y refinados.

Por ese motivo los data lakes suelen necesitar capacidades de almacenamiento mucho mayores que los almacenes de datos. Además, los datos

en bruto, sin procesar, son maleables, pueden analizarse rápidamente a cualquier efecto y son idóneos para el machine learning. El riesgo de los datos en bruto, sin embargo, es que en ocasiones estos "lagos" que son los data lakes se convierten en pantanos de datos sin la presencia de la calidad de datos y las medidas de gobernanza adecuadas.

Los almacenes de datos, como tan solo guardan datos procesados, ahorran en espacio de almacenamiento, que es un recurso caro, porque no tienen que mantener datos que quizá nunca vayan a utilizarse. Además, los datos procesados los puede entender fácilmente un público más amplio.

- Finalidad: Indeterminada o en uso

La finalidad de los componentes de datos individuales de un data lake no está establecida. Los datos en bruto se incorporan a un data lake, a veces con un uso futuro prefigurado y otras tan solo para tenerlos a mano. Esto significa que los data lakes presentan una menor organización y menor filtrado de sus datos que su equivalente.

Los datos procesados son datos en bruto a los que se ha asignado un uso concreto. Dado que los almacenes de datos tan solo albergan datos procesados, todos los datos de cualquier almacén de datos han sido utilizados para una finalidad específica dentro de la organización. Esto implica que el espacio de almacenamiento no se desperdicia en datos para los que puede que no se encuentre jamás una utilidad.

- Usuarios: Científicos de datos frente a profesionales corporativos

Para un usuario que no esté familiarizado con los datos sin procesar, los data lakes son entornos en los que cuesta orientarse. Para entender los datos en bruto no estructurados y traducirlos a una aplicación comercial específica, se necesita ser un científico de datos y contar con herramientas especializadas.

De lo contrario, cada vez son más habituales las herramientas de preparación de datos que generan un acceso en autoservicio a la información almacenada en los data lakes.

Los datos procesados se emplean en gráficos, hojas de cálculo, tablas y demás representaciones para que la inmensa mayoría de los empleados de una empresa pueda consultarlos. Los datos procesados, al igual que

los que encontramos en los almacenes de datos, tan solo exigen que el usuario tenga conocimientos de la temática representada.

- **Accesibilidad:** Flexible frente a seguro










La accesibilidad y la facilidad de uso se refieren al uso del repositorio de datos en su totalidad, no a los datos que contiene. Los data lakes no tienen estructura y, por consiguiente, resulta fácil acceder a ellos y modificarlos. Además, cualquier cambio de los datos puede efectuarse rápidamente, dado que los data lakes tienen muy pocas limitaciones.

Por diseño, los almacenes de datos son más estructurados. Una de las ventajas más destacadas de los almacenes de datos es que el procesamiento y la estructura de los datos facilita su decodificación, pero las limitaciones de su estructura complican y encarecen su manipulación.

4.8. DATA LAKE VS. DATA WAREHOUSE: PROBLEMA Y SOLUCION

- **Problema**

Ya estamos corriendo un gran almacén de datos para nuestro negocio, pero ahora la dirección quiere que construyamos un lago de datos. ¿Cual es la diferencia? ¿Dónde vamos a usarlo? ¿Quién va a usar?

	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	 Weeks - Months	\$\$\$\$\$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	 Hours - Days	\$\$\$-\$		
Data Mart	Line of Business Specific Reporting & Analytics	 Minutes - Hours	\$\$\$\$\$		

- **Solución**

Un lago de datos es una ubicación centralizada para almacenar activos de datos de una organización en su forma nativa. Idealmente, estos activos de datos se recogen de todos los puntos de contacto del negocio y se almacenan con la intención de analizar en el futuro. El propósito de la captura de todos los elementos de datos de interés es asegurar que las empresas puedan utilizar para obtener una ventaja competitiva de mercado. Desde el inicio de la informática moderna, las bases de datos se han utilizado para este propósito. lagos de datos son una extensión natural de bases de datos - y almacenes de datos posteriores - en base a la variedad de datos y cómo se almacena o se utiliza.

de dejar que un decir una empresa decide capturar información acerca de sus interacciones con el cliente. Este ha sido el papel de CRM (Customer Relationship Management) aplicaciones durante mucho tiempo. Los usuarios registrarían de ventas potenciales, la retroalimentación del cliente y otra información en una base de datos CRM. Típicamente, la base de datos sería relacional con tablas predefinidas que representan al cliente y entidades asociadas.

Las empresas modernas interactúan con los clientes de muchas maneras diferentes sin embargo: puede haber un centro de llamadas de ladrillo y mortero utilizando un CRM; pero entonces habrá uno o más sitios web con sus formularios de comentarios, correos electrónicos directos, tienda de comercio electrónico, aplicaciones móviles, presencia en medios sociales o canales de socios de negocios. Todos estos son valiosas fuentes de información que pueden proporcionar una visión 360 del cliente. Para mantener una ventaja competitiva en el mercado, la empresa tendrá que capturar información de todos estos puntos de venta.

Pero no todas las piezas de información se puede guardar en una base de datos. Algunos datos serían muy poco estructuradas, como las imágenes (piensa en los usuarios que envían imágenes de productos defectuosos); algunos pueden ser semi-estructuradas, como las redes sociales se alimenta o documentos XML. Es imposible para almacenar todo tipo de datos en una sola base de datos y es ahí donde un lago de datos puede ayudar.

En su forma más básica, un lago de datos no es más que una enorme piscina de almacenamiento donde los datos pueden ser guardados en su forma nativa, sin procesar , sin ninguna transformación aplicada.

Por ejemplo, un lago de datos puede almacenar ambas tandas nocturnas de archivos CSV (que se estructura de datos) descargadas desde el CRM y el streaming se alimenta desde el canal de medios sociales. El mismo lago de datos podría ser el hospedaje de archivos de la encuesta de satisfacción del cliente semi-estructurados enviados por terceros proveedores.

4.9. SECCION 3

4.10. SECCION 4

5. Conclusion

- Conclusion 1 :

- Conclusion 2 :

- Conclusion 3 :

Referencias

- [1] AWS (ne). What is a data lake? Recuperado de <https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/>. Accedido 20-06-2019.
 - [2] Campbell, C. (Enero 26, 2015). Top five differences between data lakes and data warehouses. Recuperado de <https://www.bluegranite.com/blog/bid/402596/top-five-differences-between-data-lakes-and-data-warehouses>. Accedido 21-06-2019.
 - [3] ColombiaDigital, C. (Junio 30, 2017). Qué es un data warehouse y qué beneficios aporta a las organizaciones? Recuperado de <https://colombiadigital.net/actualidad/articulos-informativos/item/9814-que-es-un-data-warehouse-y-que-beneficios-aporta-a-las-organizaciones.html>. Accedido 21-06-2019.
 - [4] Pearlman, S. (enero 29, 2019). Data lakes frente a almacenes de datos. Recuperado de <https://es.talend.com/resources/data-lake-vs-data-warehouse/>. Accedido 22-06-2019.
 - [5] PowerData (2015). Data lake: definición, conceptos clave y mejores prácticas. Recuperado de <https://www.powerdata.es/data-lake>. Accedido 19-06-2019.
 - [6] Sinnexus (2007). Datawarehouse. Recuperado de https://www.sinnexus.com/business_intelligence/datawarehouse.aspx. Accedido 19-06-2019.
 - [7] Soto, J. M. (Diciembre 20, 2017). Principales diferencias entre data lakes y data warehouse. Recuperado de <https://trends.inycom.es/principales-diferencias-data-lakes-data-warehouse/#prettyPhoto>. Accedido 21-06-2019.
- [1] [5] [6] [3] [4] [4] [2] [7]