Informe Tecnico, clase 21 de octubre de 2023

Daniel Soto Vasquez[1], Daniela Rodriguez Fonseca[2], Nicolás Rivas Díaz[3]

Facultad de Ingeniería Y Ciencias Básicas,
Universidad Central
Maestría en Analítica de Datos
Automatización E Integración de Datos Para IA
Bogotá, Colombia
$[1]$ dsotov1@ucentral.edu.co, [2] drodriguezf7@ucentral.edu.co [3] nrivasd@ucentral.edu.co

October 28, 2023

1 Resumen o resumen ejecutivo

El presente informe técnico se centra en la introducción a la integración de datos en donde se abordaron temas relevantes explorando aspectos, ejemplos y situaciones que permiten entender la importancia de los integracion de datos en el mundo y en las organizaciones globales.

Contents

2 Introducción

La integración de datos desempeña un papel fundamental en la eficacia de los sistemas de machine learning. Este informe casos de estudio y desarrollos recientes en el campo de la automatización e integración de datos para la IA, destacando su importancia y sus implicaciones éticas.

3 Contexto y antecedentes

La socialización de noticias, que inicia cada sesión, proporciona a los participantes en la clase de automatización e integración de datos de la Maestría en Analítica de Datos un contexto actualizado sobre avances en tecnología e inteligencia artificial. Entre las noticias destacadas se encuentra el logro de Luke Farritor al desarrollar un algoritmo de aprendizaje automático capaz de revelar letras griegas en papiros antiguos, lo que subraya la continua refinación de los procesos algorítmicos. Además, la expansión de Amazon en Europa mediante pruebas piloto de entregas con drones plantea desafíos de seguridad. También se discute la aplicación de la inteligencia artificial en la creación de modelos de riesgo por parte de empresas aseguradoras, lo que suscita preocupaciones sobre la privacidad de datos. Un estudio en una universidad de Berlín que relaciona la implementación de IA en el entorno laboral con una posible "holganazería social" destaca la necesidad de comprender los impactos sociales de estas tecnologías. Finalmente, la implementación de cámaras para el reconocimiento facial genera debates sobre la privacidad de los datos y su uso.

La exposición realizada en la clase tiene como objetivo introducir a los estudiantes al campo de la integración de datos. Se define la integración de datos como un proceso estratégico que busca unificar información procedente de diversas fuentes y sistemas para crear una vista coherente y única. Esto se presenta como esencial en un entorno empresarial impulsado por la información, ya que ayuda a superar las barreras creadas por la fragmentación de datos, permitiendo la toma de decisiones más informadas, la mejora de la eficiencia operativa y la satisfacción del cliente. Se presentan casos de estudio que demuestran cómo la integración de datos ha tenido un impacto positivo en la optimización de la cadena de suministro, la personalización de la experiencia del cliente y el análisis del rendimiento financiero en diversas industrias.

La comprensión de las fuentes de datos es crucial para la integración de datos. Se mencionan diversos tipos de fuentes de datos, como bases de datos relacionales, archivos y documentos, aplicaciones empresariales, sensores y dispositivos IoT, redes sociales y medios digitales, servicios web y API, y datos geoespaciales. Estos tipos de fuentes son fundamentales para la construcción de un sistema de integración de datos efectivo.

Además, se aborda el proceso ETL (Extract, Transform, Load), que es un componente esencial de la integración de datos. Este proceso implica la recopi-

lación de datos desde múltiples fuentes, su transformación para asegurar que sigan un formato común y su carga en un almacén de datos centralizado. Cada etapa del proceso ETL es vital para garantizar la calidad y coherencia de los datos.

4 Metodología

La metodología utilizada en la clase se muetra a continuación:

- Socialización y discusión de la bitacora de noticias
- Exposiciones individuales
- Introducción a la integración de datos

  En esta sesión no se realizó ningún laboratorio práctico.

5 Desarrollo

5.1 Socialización de noticias

Al iniciar la sesión se comienza con la socialización de noticias de la semana sobre novedades en el campo de la tecnología e inteligencia artificial. A continuación, se relacionan algunas de las noticias discutidas.

- Luke Farritor, ganó un concurso al desarrollar un algoritmo de aprendizaje automático que reveló letras griegas en papiros antiguos; demostrando así que cada vez es más refinado el proceso de procesamiento de los algoritmos.
- Por otra parte, amazon quiere expandirse a Europa, en su proceso de entrega con drones. Para lo cual van a empezar con pruebas pilotos, con el fin de mejorar su tema de seguridad.
- Existe una empresa aseguradora que entrega seguros para empresas de transportes y utilizan la inteligencia artificial para crear modelos de riesgos, ejemplo qué tan probable es que ocurra un evento (validan el comportamiento del conductor). En este punto, se discute acerca de la privacidad de los datos y cómo es manejado por estas empresas.
- Un estudio en una universidad de Berlín encuentra que las personas se sienten más relajada cuando en sus trabajos se ha implementado una IA y en consecuencia han aumentado la tasa de errores. Aquí, se discute si la implementación de IA puede causar la aparición de lo que se podría llamar "holganazería social".
- En otras noticias, hay empresas que implementan cámaras en todo lado para el reconocimiento facial. En esta noticia se discute acerca del tema de privacidad de datos de las personas, y se concluye que no debería funcionar hasta que se sepa para que están utilizando la información, cómo están guardando y tratando los datos.

5.2 Exposición

Fig1: Apache Hive - adanic

Apache Hive es una plataforma de data Warehousing construida sobre hadoop, esta permite consultar y gestionar grandes conjuntos de datos distribuidos mediante un lenguajes similar a SQL, llamado HiveSQL, esta herramienta permite análisis de datos estructurados almacenados, esto haciendo que sea muchos mas accesible a los ususarios.

5.3 Introducción a la integración de datos

¿Qué es la Integración de Datos?

La integración de datos se define como el proceso estratégico de unificar información procedente de diversas fuentes y sistemas en una vista única y coherente. En su esencia, se trata de un acto de sincronización que permite a las organizaciones sacar el máximo provecho de sus datos, convirtiéndolos en un recurso valioso y accesible.

En el ámbito empresarial, el cual actualmente es impulsado por la información, la integración de datos desempeña un papel fundamental. Permite superar las barreras impuestas por la fragmentación de datos, fomentando la toma de decisiones más informadas, optimizando la eficiencia operativa y mejorando la experiencia del cliente.

A continuación, se presentan casos de estudio en donde la implementación de integración de datos para los diferentes sistemas de información de una organización han permitido obtener beneficios no solo para la empresa sino también para los clientes.

- Optimización de la Cadena de Suministro

  Dentro de esta categoría se encuentran empresas como Walmart o Amazon, las cuales lograron cumplir la promesa de tiempo haciendo que la logística fuera definida en términos de tiempo y accesibilidad, no interesaba solo que el producto llegara al cliente o estuviera disponible en algún momento.

  * Walmart: es una grande superficie en EEUU. Después de realizar un proceso de integración de datos de las fuentes de datos; logró identificar que productos se iban acabando, lo cual permitia actualizar y adaptar toda la cadena de suministros para satisfacer la demanda. Un cliente que va a Walmart casi nunca encuetra que el producto que busca no está disponible.

* Amazon: en este caso, conectar las diferentes fases de integración de datos de todos los actores a lo largo de su cadena de suministros, la cual integra procesos y actores; permitió que varios actores funcionarán de manera coordinada; lo cual mejoro sus tiempos de entrega y su promesa de valor.

* Zara: una compañía de textiles que depende del consumo de sus productos, logró tener la moda adecuada para cada uno de los sitios en los que se encuentra. Lograron recopilar la información de ventas en tiempo real y de esta manera tomar decisiones a la hora de diseñar, producir y distribuir sus colecciones.

- Personalización de la Experiencia del Cliente

  En esta categoría se encuentran empresas como Netflix y Spotify, quienes lograron personalizar la experiencia de cada usuario en cuanto a sus gustos de películas y música respesctivamente. Al integrar sus fuentes de información, lograron los historiales de cada una de las personas para de esta forma ofrecer recomendaciones personalizadas basadas en sus gustos, lo que como consecuencia se ha logrado una mayor retención de los clientes.

- Análisis de Rendimiento Financiero

  En está categoría se encuentran algunos actores del sistema bancario como el Banco de América y la General Electric.

  Es sabido que para los bancos lo más importantes pueden ser temas como el perfilamiento de los clientes, el análisis de riesgos, temas de campañas (mercadeo) para ofrecer nuevos productos. Y para cada uno de estos temas se tienen distintas fuentes de información, ya sean manejados por la misma empresas o varios actores a lo largo de todo el sistema financiero. Y aunque, el sector bancario presenta muchas integraciones y automatizaciones, se evidencia que aun hay muchas operaciones manuales que no permiten tener una eficiencia al momento de realizar algunas transacciones como la solicitud de un crédito

5.4 Fuentes de datos

Para realizar la integración de datos se deben conocer los diferentes tipos de fuentes de datos para así conocer las diferentes herramientas de integración de datos que pueden ser utilizadas.

- Bases de Datos Relacionales: Estos almacenan datos en tablas estructuradas y se utilizan ampliamente en aplicaciones empresariales y sistemas de gestión.

- Archivos y Documentos: Datos almacenados en formatos de archivos como Excel, CSV, PDF, y documentos de texto, que contienen información valiosa.

- Aplicaciones Empresariales: Sistemas de gestión empresarial (ERP), CRM y software personalizado generan datos críticos para las operaciones y estrategias empresariales.
- Sensores y Dispositivos loT: La proliferación de sensores y dispositivos de Internet de las Cosas proporciona una gran cantidad de datos en tiempo real
- Redes Sociales y Medios Digitales: Plataformal como Facebook, Twitter, y Linkedin son fuentes ricas en datos de usuarios y tendencias.
- Servicios Web y API: Datos disponibles a través de servicios web y API, que permiten la integración con sistemas de terceros.
- Datos Geoespaciales: Información geográfica, como mapas y coordenadas, utilizada en aplicaciones de localización y logística.

5.5 Desafíos

Dentro de la integración de datos se presentan diferentes desafíos para su implementación que van desde silos de datos hasta la calidad de los datos. Esto se debe a que hay muchas variaciones posibles en la integración de datos; es decir, no hay una sola solución de integración disponible. Sin embargo, hay aspectos comunes como un servidor desde donde se extraerán los datos, fuentes dispares de datos y acceso a datos desde el servidor.

A continuación, se listan algunos de los desafíos encontrados en la integración de datos:

- Diversidad de fuentes con diferentes formatos y estructuras.
- Calidad y consistencia de los datos entre diferentes conjuntos de datos.
- Volumen de datos a integrar.
- Seguridad y Privacidad: El proceso de integración de datos y las herramientas utilizadas deben cumplir con regulaciones de privacidad y mantener la seguridad de la información sensible.
- Tiempo y costo para realizar la implementación de la integración de datos y puesta en producción.
- Documentación y Metadatos: La falta de documentación adecuada y metadatos puede dificultar la comprensión de los datos y la gestión de cambios.
- Dificultad para el uso de herramientas: No contar con personal con experiencia para implementar las plataformas de integración de datos.
- Problemas de semántica de datos: Es posible organizar varias versiones de datos que significan lo mismo o darles formato de forma distinta. Por ejemplo, las fechas que pueden ser almacenadas como número, dd/mm/aa o aaaa-mm-dd.

5.6 ETL

Los procesos ETL permiten a las organizaciones manejar flujos masivos de datos, podemos aplicar reglas y transformaciones complejas, este proceso consiste en combinar datos de diferentes orígenes un gran repositorio central llamado almacenamiento de datos. ETL utiliza un conjunto de reglas comerciales para limpiar y organizar datos en bruto y prepararlos para el almacenamiento, el análisis de datos y el machine learning (ML). Dentro del ETL se identifican 3 etapas expuestas en la siguiente imagen:

Fig1: Procesos ETL - istockphoto.com

- Extracción: En esta etapa, se recopilan datos de múltiples fuentes, que pueden incluir bases de datos, aplicaciones, archivos y sistemas externos. La extracción se realiza de manera programada y automática, asegurando que se capturen datos actualizados.
- Transformación: Una vez que los datos se han extraído, es necesario transformarlos para que se ajusten a un formato y estructura comunes. Esto puede implicar la limpieza de datos, la eliminación de duplicados, la normalización y la aplicación de reglas específicas.
- Carga (Load): Los datos transformados se cargan en un almacén de datos centralizado o en una base de datos destinada a la integración. Aquí, se pueden realizar operaciones adicionales, como la indexación, para mejorar la eficiencia de acceso.

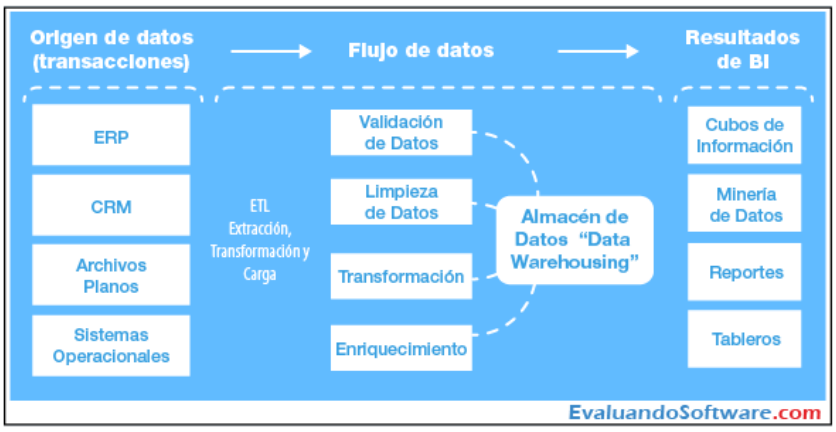

Fig2: Ejemplo ETL - evaluandosoftware.com

- En la figura 2 podemos observar un ejemplo practico de la funcionalidad ETL, en la etapa de extracción se pueden obtener datos de diversos sitemas (CRM, ERP, .dat,.csv, etc) estos pueden ser datos estructurados y no estructurados, en la etapa de transformacion se pueden aplicar diversas técnicas de valdidación, limpieza y transformación de los datos con el fin de mejorar la calidad de la misma y finalmente lograr obtener cubos de datos, hacer minería de datos, reportes y tableros de negocio, todo esto permite generar organizaciones DataDriven es decir que toman decisiones basadas en datos.

6 Resultados y análisis

Debido a que la sesión fue teórica se analiza los efectos positivos que trae realizar integración de datos en una organización. La integración de datos ha llevado beneficios significativos en diversas industrias. Sin embargo,

7 Conclusiones

- La socialización de noticias sobre tecnología e inteligencia artificial al inicio de la sesión resalta la continua evolución de la IA. Un ejemplo notable es el algoritmo desarrollado por Luke Farritor que revela letras griegas en papiros antiguos, lo que demuestra la sofisticación de los algoritmos de aprendizaje automático. Sin embargo, se deben abordar desafíos éticos, como la privacidad de datos y la seguridad, a medida que la IA se expande en diferentes industrias. La implementación de IA puede mejorar la eficiencia, pero también plantea cuestiones como la "holganazería social". En consecuencia, se destaca la necesidad de un análisis detenido y una regulación adecuada para garantizar que la IA beneficie a la sociedad.

- La exposición sobre fuentes de datos resalta la diversidad de fuentes, desde bases de datos relacionales hasta datos geoespaciales. Esta diversidad presenta un desafío en la integración de datos, ya que cada fuente puede tener su formato y estructura. El proceso ETL (Extract, Transform, Load) se identifica como esencial para combinar datos de diferentes fuentes y prepararlos para el análisis. La conclusión es que, aunque la integración de datos ofrece innumerables beneficios, los desafíos asociados, como la variedad de fuentes y la limpieza de datos, deben abordarse cuidadosamente.

- La implementación de IA en diferentes sectores ha traído notables beneficios; sin embargo, no debe olvidarse, los desafíos éticos relacionados con la privacidad y la seguridad de los datos personales y el impacto en la sociedad como la "holganazería social ", lo que requiere un análisis cuidadoso.

- La integración de datos desempeña un papel fundamental en el éxito de los proyectos de IA. La automatización y la integración de diferentes fuentes de información permiten que se pueda tomar decisiones informadas basadas en los datos casi en tiempo real. Sin embargo, es esencial abordar cuestiones éticas y de privacidad relacionadas con la recopilación y el uso de datos.

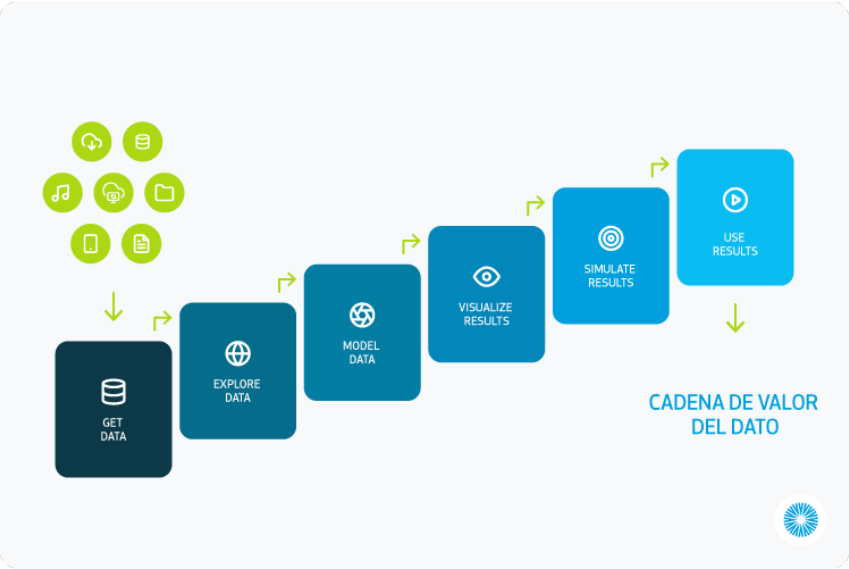

Fig3: Cadena de valor del dato

- La cadena de valor del dato, impulsada por procesos ETL eficientes, potencia la toma de decisiones informada en las organizaciones al transformar datos crudos en reales insights estratégicos, generando así un impacto directo en la eficacia operativa y la ventaja competitiva.

8 Referencias

References

[1] Amazon Web Services (AWS). (s.f.). ¿Qué es ETL (Extract, Transform, Load)? AWS. Recuperado de https://aws.amazon.com/es/what-is/etl/

[2] Google Cloud. (s.f.). ¿Qué es la integración de datos? Google Cloud. Recuperado de https://cloud.google.com/learn/what-is-data-integration?hl=es-419: :text=La

[3] Google Docs. (s.f.). Presentación de Google Docs. Recuperado de https://docs.google.com/presentation/d/1PhMUdM$_A$7$YgvWZxISGG8jfkfaMskQX8a1SOX2m0qJ8/editslide = id.g1e8906556f0_{00}$