LISTA DE COTEJO PARA INVESTIGACION

INTITUTO TECNOLOGICO SUPERIOR DE SAN ANDRES TUXTLA CARRERA: INGENIERIA INFORMÁTICA		
DATOS GENERALES DEL PROCESO DE EVALUACION		
Nombre(s) del alumno(s): **MEZO BUSTAMANTE RICARDO**		Firma del alumno(s):
Producto: Investigación Unidad I	Nombre del Proyecto: Investigar (Tecnología de Generación y Captura de Datos)	Fecha: 25/ Febrero /2023
Asignatura: **TECNICAS DE ANALISIS, MINERIA Y VISUALIZACION**	Grupo: 810 - A	Semestre: OCTAVO
Nombre del Docente: MTI Lorenzo de Jesús Organista Oliveros		Firma del Docente:

INSTRUCCIONES				
Revisar las actividades que se solicitan y marque en los apartados "SI" cuando la evidencia se cumple; en caso contrario marque "NO". En la columna OBSERVACIONES indicaciones que puedan ayudar al alumno a saber cuáles son las condiciones no cumplidas, si fuese necesario.				
Valor del reactivo	*Características a cumplir (Reactivo)*	*CUMPLE*		*OBSERVACIONES*
		SI	*NO*	
1%	**Presentación.** El trabajo cumple con los requisitos de : a. Buena presentación	X		
1%	b. No tiene faltas de ortografía	X		
1%	c. Mismo formato (letra arial 14, títulos con negritas)	X		
1%	d. Misma calidad de hoja e impresión	X		
1%	e. Maneja el lenguaje técnico apropiado	X		
2%	**Introducción y Objetivo.** La introducción y el objetivo dan una idea clara del contenido del trabajo, motivando al lector a continuar con su lectura y revisión.	X		
5%	**Sustento Teórico.** Presenta un panorama general del tema a desarrollar y lo sustenta con referencias bibliográficas y cita correctamente a los autores.	X		
2%	**Desarrollo.** Sigue una metodología y sustenta todos los pasos que se realizaron al aplicar los conocimientos obtenidos, es analítico y bien ordenado.	X		
2%	**Resultados.** Cumplió totalmente con el objetivo esperado, tiene aplicaciones concretas.	X		
2%	**Conclusiones.** Las conclusiones son claras y acordes con el objetivo esperado.	X		
2%	**Responsabilidad.** Entregó el reporte en la fecha y hora señalada.	X		
20%	**CALIFICACIÓN:**	**20%**		

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico Superior de San Andrés Tuxtla

Ingeniería Informática

Asignatura: Técnicas de Análisis, Minería y Visualización

Semestre: Octavo Semestre

Grupo: IINF810A

Docente: MTI. Lorenzo de Jesús Organista Oliveros

Alumno: Ricardo Mezo Bustamante

-Presenta-

Investigación Unidad I

Período Escolar: Febrero - Junio 2023

San Andrés Tuxtla Ver, 25 de Febrero de 2023

Introducción

En el entorno empresarial actual, la transformación digital ya no es una opción, es una necesidad. Las empresas deben adoptar nuevas tecnologías que les permitan trabajar de manera más eficiente y efectiva para ser competitivas. La captura de datos y la automatización son dos componentes críticos de la transformación digital.

Actualmente, se están desarrollando numerosas aplicaciones sociales que dan como resultado un aumento masivo **de datos** todos los días y cuando hablamos de plataformas de redes sociales, millones de usuarios se conectan a diario, la información se comparte cada vez que los usuarios usan una plataforma de redes sociales o cualquier otro sitio web. entonces surge la pregunta de cómo se maneja esta gran cantidad de datos y a través de qué medios o herramientas se procesan y almacenan los datos.

Desde la invención de las computadoras, la gente ha usado el término datos para referirse a la información de la computadora, y esta información se transmitía o almacenaba. Los datos pueden ser textos o números escritos en papeles, o pueden ser bytes y bits dentro de la memoria de dispositivos electrónicos, o pueden ser hechos almacenados dentro de la mente de una persona.

Los datos son una colección de información reunida por observaciones, mediciones, investigación o análisis. Pueden consistir en hechos, números, nombres, figuras o incluso descripción de cosas. Los datos se organizan en forma de gráficos , cuadros o tablas. Existen científicos de datos que hacen minería de datos y con la ayuda de esos datos analizan nuestro mundo.

Debe haber visto los informes del pronóstico del tiempo en los canales de noticias. Enumeran la temperatura mínima, la temperatura máxima, las predicciones y mediciones de lluvia

Las empresas confían en los datos para tomar decisiones informadas, mejorar la eficiencia operativa, comprender a sus clientes y mejorar sus productos y servicios.

Sin embargo, muchos datos todavía están atrapados en documentos y formularios físicos. Alrededor del 80% de los datos no están estructurados, lo que significa que no están en un formato que las computadoras puedan analizar fácilmente. Esto incluye datos como correos electrónicos, imágenes, archivos PDF y documentos escritos a mano. Las empresas necesitan capturar estos datos y convertirlos a un formato digital para desbloquear su valor.

Los datos están a nuestro alrededor. Se estima que cada día se crean la asombrosa cifra de 2,5 quintillones de bytes de datos. Con un número tan alucinante, es fácil ver cuánto impacto pueden tener los datos.

Con los datos apoderándose del mundo, es posible que tenga la curiosidad de aprender un poco sobre la ciencia de datos. Aquí es donde entra la captura de datos.

Datos

Los datos son diferentes tipos de información generalmente formateada de una manera particular. Todo el software se divide en dos categorías principales: programas y datos. Ya sabemos qué son los datos ahora, y los programas son colecciones de instrucciones que se utilizan para manipular datos.

Utilizamos la ciencia de datos para facilitar el trabajo con datos. La ciencia de datos se define como un campo que combina el conocimiento de las matemáticas, las habilidades de programación , la experiencia en el dominio, los métodos científicos, los algoritmos, los procesos y los sistemas para extraer conocimientos e ideas procesables de datos estructurados y no estructurados, y luego aplicar el conocimiento obtenido de esos datos para una amplia gama de usos y dominios.

En computación, los datos son información que ha sido traducida a una forma que es eficiente para su movimiento o procesamiento. En relación con las computadoras y los medios de transmisión actuales, los datos son información convertida en forma digital binaria . Es aceptable que los datos se utilicen como un sujeto singular o un sujeto plural. Raw data es un término utilizado para describir datos en su formato digital más básico.

El concepto de datos en el contexto de la computación tiene sus raíces en el trabajo de Claude Shannon, un matemático estadounidense conocido como el padre de la teoría de la información. Dio comienzo a los conceptos digitales binarios basados en la aplicación de la lógica booleana de dos valores a los circuitos electrónicos. Los formatos de dígitos binarios son la base de las CPU , las memorias de semiconductores y las unidades de disco, así como muchos de los dispositivos periféricos comunes en la informática actual. Las primeras entradas informáticas tanto para el control como para los datos tomaron la forma de tarjetas perforadas , seguidas por la cinta magnética y el disco duro.

Al principio, la importancia de los datos en la informática comercial se hizo evidente por la popularidad de los términos "procesamiento de datos" y "procesamiento electrónico de datos", que, durante un tiempo, llegaron a abarcar toda la gama de lo que ahora se conoce como tecnología de la información . A lo largo de la historia de la informática corporativa, se produjo la especialización y surgió una profesión de datos distinta junto con el crecimiento del procesamiento de datos corporativos.

Cómo Se Almacenan Los Datos

Las computadoras representan datos, incluidos videos, imágenes, sonidos y texto, como valores binarios utilizando patrones de solo dos números: 1 y 0. Un bit es la unidad de datos más pequeña y representa solo un valor. Un byte tiene ocho dígitos binarios. El almacenamiento y la memoria se miden en megabytes y gigabytes .

Las unidades de medida de datos continúan creciendo a medida que crece la cantidad de datos recopilados y almacenados. El término relativamente nuevo

" brontobyte ", por ejemplo, es el almacenamiento de datos que es igual a 10 a la 27 potencia de bytes .

Los datos se pueden almacenar en formatos de archivo, como en los sistemas de mainframe que utilizan ISAM y VSAM . Otros formatos de archivo para almacenamiento, conversión y procesamiento de datos incluyen valores separados por comas. Estos formatos continuaron encontrando usos en una variedad de tipos de máquinas, incluso cuando los enfoques orientados a datos más estructurados ganaron terreno en la informática corporativa.

Se desarrolló una mayor especialización como base de datos , sistema de gestión de base de datos y luego surgió la tecnología de base de datos relacional para organizar la información.

Manejo Y Uso De Datos

Con la proliferación de datos en las organizaciones, se ha puesto mayor énfasis en garantizar la calidad de los datos al reducir la duplicación y garantizar que se utilicen los registros más actualizados y precisos. Los muchos pasos involucrados con la gestión de datos moderna incluyen la limpieza de datos, así como los procesos de extracción, transformación y carga (ETL) para integrar datos. Los datos para el procesamiento se han complementado con metadatos, a veces denominados "datos sobre datos", que ayudan a los administradores y usuarios a comprender la base de datos y otros datos.

Los análisis que combinan datos estructurados y no estructurados se han vuelto útiles, ya que las organizaciones buscan capitalizar dicha información. Los sistemas para tales análisis se esfuerzan cada vez más por lograr un rendimiento en tiempo real, por lo que están diseñados para manejar los datos entrantes consumidos a altas tasas de ingesta y para procesar flujos de datos para su uso inmediato en las operaciones.

Con el tiempo, la idea de la base de datos para operaciones y transacciones se ha extendido a la base de datos para informes y análisis de datos predictivos. Un ejemplo principal es el almacén de datos , que está optimizado para procesar preguntas sobre operaciones para analistas comerciales y líderes comerciales. El creciente énfasis en la búsqueda de patrones y la predicción de los resultados comerciales ha llevado al desarrollo de técnicas de minería de datos.

¿Quién Crea Los Datos?

Los datos pueden ser creados en una computadora por el usuario , el software o el hardware conectado a la computadora. En el caso de un usuario, los datos se ingresan usando un dispositivo de entrada como un teclado . Con el software, un programa puede crear datos de los usuarios que interactúan con el programa o de otra fuente de entrada (por ejemplo, una alerta activada). Finalmente, el hardware

también puede crear datos a partir de una alerta recibida por un sensor u otra entrada.

Diferencia Entre Datos E Información.

Los datos son hechos simples. La palabra "datos" es plural para "dato". Cuando los datos se procesan, organizan, estructuran o presentan en un contexto determinado para que sean útiles, se denominan Información.

No basta con tener datos (como estadísticas sobre la economía). Los datos en sí mismos son bastante inútiles, pero cuando estos datos se interpretan y procesan para determinar su verdadero significado, se vuelven útiles y pueden denominarse Información.

La datos son información que han sido procesados de tal manera que sean significativos para la persona que los recibe. es cualquier cosa que se comunica.

Datos es el término, que puede ser nuevo para los principiantes, pero es muy interesante y fácil de entender. Puede ser cualquier cosa, como el nombre de una persona, un lugar, un número, etc. Los datos son el nombre que se le da a hechos y entidades básicos, como nombres y números. Los principales ejemplos de datos son pesos, precios, costos, cantidad de artículos vendidos, nombres de empleados, nombres de productos, direcciones, códigos de impuestos, marcas de registro, etc.

Los datos son la materia prima que puede ser procesada por cualquier máquina informática. Los datos se pueden representar en forma de: Números y palabras que se pueden almacenar en el lenguaje de la computadora, Imágenes, sonidos, multimedia y datos animados.

Información: La información son datos que se han convertido en una forma más útil o inteligible. Es el conjunto de datos que se ha organizado para la utilización directa de la humanidad, ya que la información ayuda a los seres humanos en su proceso de toma de decisiones. Algunos ejemplos son: calendario, lista de méritos, boleta de calificaciones, tablas con encabezados, documentos impresos, nóminas, recibos, informes, etc. La información se obtiene reuniendo elementos de datos en una forma significativa.

Por ejemplo, las calificaciones obtenidas por los estudiantes y sus números de lista forman datos, la boleta/hoja de calificaciones es la información. Otras formas de información son nóminas, horarios, informes, hojas de trabajo, gráficos de barras, facturas y devoluciones de cuentas, etc. Cabe señalar que la información puede procesarse y/o manipularse posteriormente para formar conocimiento. La información que contiene sabiduría se conoce como conocimiento.

Captura de Datos

La captura de datos es el proceso de recopilar información de un documento y convertirla en datos que las computadoras puedan entender. Es una de las fases más esenciales de la digitalización y, si se realiza correctamente, permitirá a los empleados almacenar, organizar, buscar y recuperar documentos en un tiempo récord.

La captura de datos se refiere a extraer datos de documentos físicos o digitales, como archivos PDF o imágenes, y convertirlos a formatos que las computadoras puedan entender. Los formatos que las computadoras pueden leer incluyen CSV (valores separados por comas), XML (lenguaje de marcado extensible) y JSON (notación de objetos de JavaScript).

El objetivo principal de la captura de datos es poder transformar la información de todas las fuentes en un formato que las computadoras puedan entender. Los datos recopilados revelarán información sobre cómo le está yendo a su organización y cómo mejorar mediante el aumento de la eficiencia.

En el pasado, la captura de datos era un proceso manual. Las empresas tendrían que contratar trabajadores para ingresar manualmente datos de documentos en papel a formatos digitales. Esto no solo requería mucho tiempo, sino que también era propenso a errores. Incluso un simple error tipográfico podría resultar en una entrada de datos incorrecta en el sistema, lo que generaría problemas posteriores.

Ahora, con tecnologías más avanzadas, las empresas pueden automatizar la captura de datos. La captura de datos automatizada utiliza tecnología, como el reconocimiento óptico de caracteres (OCR) y el reconocimiento inteligente de documentos (IDR), para extraer datos de los documentos automáticamente. Esta es una forma mucho más eficiente y precisa de capturar datos, eliminando la necesidad de una entrada manual.

Las organizaciones pueden utilizar las tecnologías de captura de datos de diversas formas, como, por ejemplo:

- Los minoristas pueden utilizar la captura de datos para extraer automáticamente los datos de los recibos . Esto se puede usar para rastrear el historial de compras del cliente y comprender mejor el comportamiento del consumidor.
- Las instituciones financieras pueden utilizar la captura de datos para automatizar la incorporación de clientes. Al capturar datos de pasaportes, identificaciones y comprobantes de domicilio , los bancos y las compañías de seguros pueden verificar rápida y fácilmente la identidad del cliente.
- Las organizaciones de atención médica pueden usar la captura de datos para optimizar la admisión de pacientes. Al capturar datos de los formularios de

los pacientes, los hospitales pueden acelerar el proceso de admisión y reducir el papeleo.

- Los gobiernos pueden utilizar la captura de datos para reducir el fraude y acelerar el procesamiento de solicitudes de beneficios. Al capturar datos de tarjetas de identificación , certificados de nacimiento y otros documentos, las agencias gubernamentales pueden verificar la identidad de los solicitantes y prevenir el fraude.
- La captura de datos también puede extraer y digitalizar datos de certificados comerciales . Los detalles como el nombre de la empresa, el número de registro y la fecha de incorporación se pueden capturar y almacenar rápidamente en un formato digital.

Como puede ver, la captura de datos tiene una amplia gama de aplicaciones. Se puede utilizar para automatizar varios procesos comerciales y flujos de trabajo, lo que puede conducir a una mayor eficiencia y productividad.

¿Cómo Capturar Datos?

Para comprender mejor cómo funciona la captura de datos, debemos analizar algunos de los diferentes métodos que utilizan las empresas. Estos son los dos tipos principales:

Captura manual de datos

Este es el enfoque tradicional para capturar datos de documentos. Como sugiere el nombre, la captura manual de datos implica ingresar información manualmente en un sistema informático. Uno tiene que ingresar lo que se captura del documento y hacer cualquier validación requerida.

Las desventajas de la captura manual de datos son obvias. En primer lugar, requiere mucho tiempo. En un momento en que las cosas se mueven rápido y el tiempo es esencial para muchas empresas, no es factible ingresar datos manualmente.

En segundo lugar, es propenso a errores de entrada de datos . Los humanos cometen errores y, como resultado, a menudo necesitan regresar y volver a ingresar datos. Esto conduce inevitablemente a pérdidas de productividad e ineficiencias operativas.

Además, la captura manual de datos no escala bien. A medida que una empresa crece y procesa más documentos, este enfoque rápidamente se vuelve insostenible.

Captura de datos automatizada

Esta es la alternativa moderna a la captura manual de datos. En lugar de ingresar manualmente la información de un documento o formulario, las empresas pueden usar software de automatización para extraer esa información automáticamente. Gracias a la inteligencia artificial (IA) y el aprendizaje

automático , este software puede comprender el diseño del documento y extraer datos relevantes con poca intervención humana.

Métodos de captura de datos

El crecimiento de la tecnología de la información ha dado como resultado que la mayoría de los datos se vuelvan digitales, como archivos de documentos, archivos PDF, formularios electrónicos, correos electrónicos, videos, etc. Sin embargo, una cantidad sustancial de datos todavía se crea manualmente, como documentos en papel, cartas y certificados de trabajo.

Existen varios métodos de captura de datos para capturar información de investigaciones, encuestas, correos electrónicos, facturas y otras fuentes. Se clasifican en dos tipos: manuales y automáticos. En la captura automatizada de datos se utilizan tecnologías avanzadas como OCR , códigos de barras, firmas digitales , procesamiento inteligente de documentos , etc.

Elegir los métodos de captura de datos inteligentes apropiados para usar es una disciplina en constante desarrollo. Para hacer esto, las empresas deben emplear una variedad de formas de captura de información. Las empresas deben identificar explícitamente las mejores prácticas y técnicas que se utilizarán en función del tipo de contenido. Debido a que no toda la información se crea de la misma manera, es posible que deba emplear una variedad de métodos de captura de datos digitales.

Hay diferentes formas de automatizar la captura de datos, que incluyen:

Reconocimiento Óptico De Caracteres (OCR)

OCR implica escanear texto dentro de una imagen o PDF y extraerlo. Luego convierte ese texto en texto codificado por máquina. Desafortunadamente, la información no está estructurada, por lo que los resultados no pueden ser utilizados por una computadora.

Reconocimiento Inteligente De Caracteres (ICR)

ICR es una versión más avanzada de OCR. Esta versión puede incluso reconocer caracteres escritos a mano y extraer datos de ellos. Sin embargo, ICR también sufre el mismo problema que OCR, ya que solo puede extraer datos no estructurados.

Procesamiento Inteligente De Documentos (IDP)

El procesamiento inteligente de documentos es el enfoque más avanzado y sofisticado para la captura de datos. Utiliza tecnologías avanzadas como inteligencia artificial, aprendizaje automático y procesamiento de lenguaje natural para comprender documentos y extraer datos de ellos.

Puede extraer automáticamente datos de imágenes, archivos PDF y documentos escritos a mano mientras comprende el contexto y lo organiza en un formato

estructurado. Esto hace posible utilizar los datos directamente con fines comerciales.

Códigos De Barras Y Códigos QR

Los códigos de barras o códigos QR se utilizan a menudo para almacenar datos sobre un producto, como su precio o número SKU. Esta información se puede capturar utilizando un escáner de código de barras e importarse a un sistema informático.

Formularios Digitales

Este método facilita la captura de datos a través de aplicaciones web y móviles. Es personalizable y elimina la necesidad de formularios en papel. Los formularios digitales ofrecen mayor seguridad y privacidad, ya que los datos se almacenan electrónicamente y pueden protegerse con contraseña. Además, se pueden integrar con sistemas back-end para una eficiencia aún mayor.

Firmas Digitales

Las firmas digitales a menudo reemplazan las firmas en papel y se pueden usar para firmar documentos, contratos y otros acuerdos legales.

Las firmas digitales son más seguras que las firmas tradicionales, ya que no se pueden falsificar. También proporcionan un registro de cuándo se firmó un documento y quién lo firmó, lo que puede ser útil para fines de auditoría.

Web Scraping

Este método implica el uso de bots y rastreadores web para buscar y recopilar datos de Internet. Estos datos pueden luego transferirse a bases de datos relevantes para su uso. La principal ventaja del web scraping es que se puede utilizar para recopilar datos que cambian constantemente, como datos meteorológicos o precios de acciones.

Captura De Voz

La captura de voz es un método de entrada de datos que implica el uso de software de reconocimiento de voz para convertir grabaciones de audio en texto. Este texto se puede transferir a un sistema informático para su posterior procesamiento.

También se puede utilizar para comprender e interpretar comandos hablados. Siri de Apple, Alexa de Amazon y Cortana de Microsoft son ejemplos de tecnología de captura de voz.

Existen muchos otros métodos de captura de datos, incluidas tarjetas de banda magnética, lectura de marcas ópticas, reconocimiento de caracteres de tinta magnética, tarjetas inteligentes, captura de video/imágenes y más. Sin embargo, estos son los métodos más comunes utilizados hoy en día.

Con los avances en la computación en la nube, la inteligencia artificial y la tecnología móvil, la captura de datos ha recorrido un largo camino en los últimos años y se requiere de alguna forma en casi todos los procesos digitales. El mundo digital puede superponerse y coexistir con el mundo físico y las operaciones comerciales para crear nuevos valores y posibilidades en nuestra vida personal y laboral.

El Futuro De La Industria De Captura De Datos Con IA

A medida que la aplicación de IA crece exponencialmente minuto a minuto, lo siguiente muestra una vista previa del futuro de la industria de captura de datos:

1. Mayor velocidad de análisis

Anteriormente, el procesamiento y análisis de datos tenía que hacerse manualmente. Estas actividades se realizan instantáneamente con tecnología de inteligencia artificial, lo que permite a las empresas resolver problemas más rápido y permite que los profesionales de gestión de datos se concentren en otras responsabilidades centrales y más vitales.

2. Análisis de transmisión en tiempo real

Las organizaciones realizarán análisis de transmisión en tiempo real para adquirir datos actualizados, precisos al segundo, a medida que la IA se integre progresivamente en la empresa.

3. Los flujos de trabajo de DevOps superarán al desarrollo de aplicaciones

Las sugerencias y recomendaciones se perfeccionarán a medida que los sistemas de inteligencia artificial y aprendizaje automático aprendan y crezcan. Este flujo de trabajo podría llevar a que más empresas implementen flujos de trabajo DevOps basados en IA para el desarrollo de aplicaciones, lo que permite a los ingenieros integrar y ofrecer actualizaciones de software que aprovechan continuamente el conocimiento y el aprendizaje de la IA.

4. La IA transformará todas las industrias

Las industrias cambiarán drásticamente a medida que la IA avance y las empresas creen flujos de trabajo que les permitan maximizar el valor. Con herramientas impulsadas por IA y otras innovaciones, los proveedores de atención médica, los bancos, las empresas de transporte, los recursos humanos y cualquier otra industria podrán brindar servicios más eficientes y rentables.

Conclusiones

Los datos son la mina de oro de la empresa. La captura de datos se ha convertido en una herramienta fundamental para impulsar a las organizaciones hacia operaciones y eficiencia mejoradas.

El futuro de la captura inteligente de datos está en la integración de tecnologías cognitivas avanzadas como IA y ML. Cuantas más capacidades de inteligencia artificial se incluyan, más excelente será la calidad de los datos extraídos.

La integración de tecnologías cognitivas sofisticadas como AI y ML es el futuro de la captura de datos cognitivos. Cuantas más capacidades de inteligencia artificial se integren, mayor será la calidad de los datos recuperados.

¡Es razonable afirmar que las tecnologías de captura de datos automatizadas realmente han transformado las organizaciones hoy en día!

La extracción de datos de conjuntos de datos masivos está transformando la forma en que pensamos sobre la respuesta a las crisis, el marketing, el entretenimiento, la ciberseguridad y la inteligencia nacional. Las colecciones de documentos, imágenes, videos y redes se consideran no solo como cadenas de bits para almacenar, indexar y recuperar, sino como fuentes potenciales de descubrimiento y conocimiento, que requieren técnicas de análisis sofisticadas que van mucho más allá de la indexación clásica y el conteo de palabras clave; con el objetivo de encontrar interpretaciones relacionales y semánticas de los fenómenos que subyacen a los datos.

La captura de datos se ha convertido en una herramienta inevitable para impulsar a las empresas hacia un mejor funcionamiento y productividad. La llegada de la IA ha mejorado la forma en que se capturan los datos para crear nuevas posibilidades. Los datos son extremadamente precisos, muy accesibles y han abierto nuevas puertas de enlace para que las empresas se aseguren de estar en la cima. ¡Es seguro decir que las tecnologías de captura de datos automatizadas realmente se han convertido en un punto de inflexión para las empresas de hoy!

Referencias Bibliográficas

[1]"¿Qué es la captura de datos y por qué la necesita? | FormX.ai". FormX.ai - Extractor de formularios y documentos con IA. https://www.formx.ai/post/what-is-data-capture-and-how-to-use-it#definition (accedido el 26 de febrero de 2023).

[2]"¿Qué es la captura de datos y por qué es importante?" Gestión de la información simplificada. https://theecmconsultant.com/what-is-data-capture/ (accedido el 26 de febrero de 2023).

[3]S.Jameson. "Captura de datos | ¿Qué es la captura de datos?" Blog sobre inteligencia artificial y aprendizaje automático sobre nanoredes. https://nanonets.com/blog/what-is-data-capture/ (accedido el 26 de febrero de 2023).

[4]"Métodos de captura de datos - ProcessFlows UK Ltd". ProcessFlows UK Ltd. https://processflows.co.uk/data-capture-digitisation/methods-of-data-capture/ (accedido el 26 de febrero de 2023).

[5]"¿Qué es la captura de datos? - Una guía completa [actualización de 2021]". Servicios de datos de iTech. https://itechdata.ai/what-is-data-capture-and-how-can-your-business-benefit-from-using-it/ (accedido el 26 de febrero de 2023).

[6]"What is Data?" Computer Hope's Free Computer Help. https://www.computerhope.com/jargon/d/data.htm#who (accedido el 26 de febrero de 2023).

[7]J. Vaughan. "¿Qué son los datos? - Definición de WhatIs.com". Gestión de datos. https://www.techtarget.com/searchdatamanagement/definition/data (accedido el 26 de febrero de 2023).

[8]"¿Cuál es la diferencia entre datos e información? - Notas de computadora". Notas de la computadora. https://ecomputernotes.com/fundamental/information-technology/what-do-you-mean-by-data-and-information (accedido el 26 de febrero de 2023).

ITSSAT

GUIA DE OBSERVACIÓN PARA RESOLUCIÒN DE EJERCICIOS PRACTICOS

NOMBRE DE LA ASIGNATURA: **TECNICAS DE ANALISIS, MINERIA Y VISUALIZACION**

NOMBRE DE LA UNIDAD: **INTRODUCCIÓN EL CICLO DE VIDA DEL DATOS**

ALUMNO: **MEZO BUSTAMANTE RICARDO**

INSTRUCCIONES

Revisar los documentos o actividades que se solicitan y marque en los apartados "SI" cuando la evidencia a evaluar se cumple; en caso contrario marque "NO". En la columna "OBSERVACIONES" ocúpela cuando tenga que hacer comentarios referentes a lo observado.

Valor del reactivo	***Características a cumplir (Reactivo)***	***CUMPLE***		***OBSERVACIONES***
		Si	***NO***	
8%	¿Identifico el problema planteado?	X		
4%	¿Identifico los datos de entrada del problema?	X		
4%	¿Identifico los datos de salida del problema?	X		
8%	¿Generó la solución del problema en forma clara y comprensible (orden)?	X		
12%	¿Elaboró el programa respetando la sintaxis del lenguaje de programación (orden)?	X		
4%	Comprobó los resultados esperados a través de pruebas de escritorio?	X		
40%	***CALIFICACIÓN:***	**40%**		

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico Superior de San Andrés Tuxtla

Ingeniería Informática

Asignatura: Técnicas de Análisis, Minería y Visualización

Semestre: Octavo Semestre

Grupo: IINF810A

Docente: MTI. Lorenzo de Jesús Organista Oliveros

Alumno: Ricardo Mezo Bustamante

-Presenta-

Practica Unidad I

Período Escolar: Febrero - Junio 2023

San Andrés Tuxtla Ver, 28 de Febrero de 2023

Introducción

En esta era de Big Data, las empresas de todos los niveles están expuestas a un volumen cada vez mayor de flujos de datos de una amplia gama de fuentes. Estos conjuntos de datos son de misión crítica para ayudarlos a optimizar sus estrategias de marketing. Sin embargo, para hacer el mejor uso de los datos para tomar decisiones productivas, las empresas deben extraer dichos datos de todas las fuentes disponibles y consolidarlos en un destino para un análisis y una gestión de datos óptimos.

Los estudios recientes revelan que las organizaciones basadas en datos tienen 23 veces más probabilidades de obtener nuevos clientes, 6 veces más probabilidades de mantener una relación con ellos y 19 veces más probabilidades de ser rentables. Esto demuestra que los datos ahora controlan el mundo de la empresa moderna**.**

Actualmente, las empresas dependen en gran medida de los datos para predecir tendencias, pronosticar el mercado, planificar las necesidades futuras, comprender a los consumidores y tomar decisiones comerciales. Sin embargo, para realizar tales tareas, es fundamental tener acceso rápido a los datos de la empresa en una ubicación centralizada. La tarea de recopilar y almacenar datos estructurados y no estructurados en una ubicación centralizada se denomina **ingesta de datos**.

Uno de los desafíos clave que enfrentan las empresas modernas es el enorme volumen de datos de numerosas fuentes de datos. Estamos en la era de Big Data, donde los datos se inundan a un ritmo sin precedentes y es difícil recopilar y procesar estos datos sin las herramientas de manejo de datos adecuadas.

Elegir la herramienta adecuada no es tarea fácil, y más difícil es manejar grandes volúmenes de datos si la empresa no conoce las herramientas disponibles. Sin embargo, muchas empresas contemporáneas que manejan cantidades sustanciales de datos utilizan diferentes tipos de herramientas para cargar y procesar datos de diversas fuentes de manera eficiente y eficaz.

La ingesta de datos es uno de los primeros pasos del proceso de manejo de datos. Con las herramientas de ingestión de datos adecuadas, las empresas pueden recopilar, importar, procesar y almacenar rápidamente datos de diferentes fuentes de datos.

La inteligencia que impulsa el análisis en tiempo real, las aplicaciones inteligentes y las operaciones de aprendizaje automático comienza con los datos. ¡Montones y montones de datos! Obtener datos de todas partes donde el equipo de datos pueda usarlos para la innovación y el crecimiento comienza con la ingestión de datos.

Ingesta de Datos

La ingesta de datos implica ensamblar datos de varias fuentes en diferentes formatos y cargarlos en un almacenamiento centralizado, como un lago de datos o un almacén de datos. Se accede a los datos almacenados y se analizan para facilitar las decisiones basadas en datos. Los sistemas de procesamiento de datos pueden incluir lagos de datos, bases de datos y repositorios de almacenamiento dedicados. Al implementar la ingesta de datos, los datos pueden ingerirse en lotes o transmitirse en tiempo real. Cuando los datos se incorporan en lotes, se importan en fragmentos discretos a intervalos regulares, mientras que en la incorporación de datos en tiempo real, cada elemento de datos se importa continuamente a medida que lo emite la fuente.

La ingestión de datos es el proceso de mover datos de una fuente a un área de destino o un almacén de objetos donde se pueden usar para consultas y análisis ad hoc. Una canalización de ingesta de datos simple consume datos desde un punto de origen, los limpia un poco y luego los escribe en un destino.

La ingestión de datos es el proceso de importar archivos de datos grandes y variados de varias fuentes a un único medio de almacenamiento basado en la nube (un almacén de datos, un data mart o una base de datos) donde se puede acceder a ellos y analizarlos. Como los datos pueden estar en múltiples formas diferentes y provenir de cientos de fuentes, se desinfectan y transforman en un formato uniforme mediante un proceso de extracción/transformación/carga (ETL).

La ingestión de datos es el proceso de transporte de datos desde una o más fuentes a un sitio de destino para su posterior procesamiento y análisis. Estos datos pueden provenir de una variedad de fuentes, incluidos lagos de datos, dispositivos IoT, bases de datos locales y aplicaciones SaaS, y terminar en diferentes entornos de destino, como almacenes de datos en la nube o data marts.

La ingestión de datos es una tecnología crítica que ayuda a las organizaciones a dar sentido a un volumen y una complejidad de datos cada vez mayores. Para ayudar a las empresas a obtener más valor de la ingestión de datos, profundizaremos en esta tecnología. Cubriremos los tipos de ingesta de datos, cómo se realiza la ingesta de datos, la diferencia entre la ingesta de datos y ETL, las herramientas de ingesta de datos y más.

La ingesta de datos es el proceso de importar y cargar datos en un sistema. Es uno de los pasos más críticos en cualquier flujo de trabajo de análisis de datos. Una empresa debe ingerir datos de varias fuentes, incluidas las plataformas de marketing por correo electrónico, los sistemas de CRM, los sistemas financieros y las plataformas de redes sociales. Los científicos de datos suelen realizar la ingestión de datos porque requiere experiencia en aprendizaje automático y lenguajes de programación como Python y R.

¿Por Qué Es Tan Importante La Ingesta De Datos?

La ingestión de datos ayuda a los equipos a ir rápido. El alcance de cualquier canal de datos dado es deliberadamente estrecho, lo que brinda a los equipos de datos flexibilidad y agilidad a escala. Una vez que se establecen los parámetros, los analistas de datos y los científicos de datos pueden crear fácilmente una única canalización de datos para mover los datos al sistema de su elección. Los ejemplos comunes de ingestión de datos incluyen:

- Mueva datos de Salesforce.com a un almacén de datos y luego analícelos con Tableau
- Capture datos de un feed de Twitter para el análisis de sentimientos en tiempo real
- Adquirir datos para entrenar modelos de aprendizaje automático y experimentación

La Integración De Datos Moderna Comienza Con La Ingesta De Datos

Los ingenieros de datos utilizan canalizaciones de ingesta de datos para gestionar mejor la escala y la complejidad de las demandas empresariales de datos. Muchas canalizaciones de datos impulsadas por la intención que operan continuamente en toda la organización sin la participación directa de un equipo de desarrollo permiten una escala sin precedentes para lograr objetivos comerciales importantes. Éstas incluyen:

- Acelere los pagos para una red global de proveedores de atención médica a través de microservicios
- Apoye las innovaciones de IA y los casos de uso comercial con una plataforma de datos de autoservicio
- Descubra el fraude con la ingestión y el procesamiento en tiempo real en un lago de datos 360 del cliente

La ingestión de datos se ha convertido en un componente clave de las plataformas de autoservicio para que analistas y científicos de datos accedan a datos para análisis en tiempo real, aprendizaje automático y cargas de trabajo de IA.

Cómo Funciona La Ingesta De Datos

La ingestión de datos extrae datos del origen donde se crearon o almacenaron originalmente, y los carga en un destino o área de preparación. Una canalización de ingesta de datos simple podría aplicar una o más transformaciones ligeras que enriquecen o filtran los datos antes de escribirlos en algún conjunto de destinos, un almacén de datos o una cola de mensajes. Se pueden realizar transformaciones más complejas, como uniones, agregados y clasificaciones para análisis, aplicaciones y sistemas de informes específicos, con canalizaciones adicionales.

Herramientas de Ingesta de Datos					
Herramienta	Que Es	Características	Ventajas	Desventajas	Casos de Uso
Apache Kafka	Kafka es un sistema distribuido que consta de servidores y clientes que se comunican a través de un protocolo de red TCP de alto rendimiento. Se puede implementar en hardware básico, máquinas virtuales y contenedores en entornos locales y en la nube.	**1. Escalabilidad**: Apache Kafka puede manejar la escalabilidad en las cuatro dimensiones, es decir, productores de eventos, procesadores de eventos, consumidores de eventos y conectores de eventos. En otras palabras, Kafka escala fácilmente sin tiempo de inactividad. **2. Alto volumen:** Apache Kafka puede trabajar fácilmente con un gran volumen de flujos de datos. **3. Transformaciones de datos**: Apache Kafka ofrece disposiciones para derivar nuevos flujos de datos utilizando los flujos de datos de los productores. **4. Tolerancia a fallas**: Los clústeres de	Kafka fue diseñado para ofrecer estas ventajas distintivas sobre AMQP, JMS, etc. 1. **Kafka es altamente escalable.** Kafka es un sistema distribuido, que se puede escalar rápida y fácilmente sin incurrir en ningún tiempo de inactividad. Apache Kafka es capaz de manejar muchos terabytes de datos sin incurrir en muchos gastos generales. 2. **Kafka es muy duradero.** Kafka conserva los mensajes en los discos, lo que proporciona replicación dentro del clúster. Esto lo convierte en un sistema de mensajería muy duradero. 3. **Kafka es altamente confiable.** Kafka replica datos y puede	1. No posee un conjunto completo de herramientas de administración y monitoreo. 2. El corredor usa ciertas llamadas al sistema para entregar mensajes al consumidor, pero si el mensaje necesita algunos ajustes, hacerlo reduce significativamente el rendimiento de Kafka.	Kafka se ha utilizado ampliamente, y es una parte integral de la pila de Spotify, Netflix, Uber, Goldman Sachs, Paypal, etc., que lo utilizan para procesar datos de streaming y comprender el comportamiento de los clientes o del sistema. En realidad, Kafka ha ganado dominio en la industria de los viajes, donde su capacidad de streaming lo hace ideal para el seguimiento de los detalles de reservas de millones de vuelos, paquetes de vacaciones y vacantes de hotel en todo el mundo.

		Apache Kafka pueden manejar fallas con los maestros y las bases de datos. **5. Confiabilidad**: Dado que Apache Kafka está distribuido, particionado, replicado y tolerante a fallas, es muy confiable. **6. Durabilidad**: Apache Kafka es duradero porque utiliza registros de confirmación distribuidos, lo que significa que los mensajes persisten en el disco lo más rápido posible. **7. Rendimiento**: Tanto para publicar como para suscribirse a mensajes, Kafka tiene un alto rendimiento. Incluso si se almacenan muchos TB de mensajes, mantiene un rendimiento estable.	admitir múltiples suscriptores. Además, equilibra automáticamente a los consumidores en caso de falla. Eso significa que es más confiable que los servicios de mensajería similares disponibles. 4. **Kafka ofrece alto rendimiento.** Kafka ofrece un alto rendimiento tanto para la publicación como para la suscripción, utilizando estructuras de disco capaces de ofrecer niveles constantes de rendimiento, incluso cuando se trata de muchos terabytes de mensajes almacenados.		

		8. Cero tiempo de inactividad: Apache Kafka es muy rápido y garantiza cero tiempo de inactividad y cero pérdida de datos. **9. Replicación**: Kafka MirrorMaker brinda soporte de replicación para sus clústeres. Con funciones de replicación, los mensajes se replican en varios centros de datos o regiones de la nube. Puede usar estos escenarios inactivos/pasivos para respaldo y recuperación, o escenarios inactivos/activos para colocar los datos más cerca de sus usuarios o cumplir con los requisitos de localidad de datos. **10. Gratis para usar**: Apache Kafka fue desarrollado			

		por LinkedIn y posteriormente donado a Apache Software Foundation. ¡Debido a que es de código abierto, no hay tarifas de licencia para usar Kafka! Este software es completamente gratuito. Estas son las características principales de Apache Kafka que hacen de Kafka una poderosa herramienta para administrar el flujo de datos en tiempo real.			
Apache Sqoop	Apache Sqoop(TM) es una herramienta diseñada para transferir de manera eficiente datos masivos entre Apache Hadoop y almacenes de datos estructurados, como bases de datos relacionales. Apache Sqoop es parte del ecosistema Hadoop.	Apache Sqoop tiene muchas características esenciales. Algunos de ellos se discuten aquí: • Sqoop utiliza el marco YARN para importar y exportar datos. El paralelismo se ve reforzado por la tolerancia a fallas de esta manera.	Estas son algunas de las ventajas de Apache Sqoop que hacen que este importante aspecto del ecosistema de Hadoop sea tan popular. • Implica transferir datos de una variedad de fuentes estructuradas de datos como Oracle, Postgres, etc.	Aunque Sqoop tiene ventajas muy fuertes en su nombre, tiene algunas desventajas inherentes, que se pueden resumir como: • Utiliza una conexión JDBC para conectarse con almacenes de datos basados en RDBMS, y esto puede ser ineficiente y de menor rendimiento.	Se puede utilizar para procesar grandes cantidades de datos genómicos y otros conjuntos de datos científicos de gran tamaño de forma rápida y eficiente. AWS ha puesto a disposición de la comunidad los datos del proyecto de los 1 000 genomas de forma gratuita. La herramienta utiliza MapReduce para realizas dichas operaciones, por lo que

	Dado que muchos de los datos debían transferirse desde los sistemas de bases de datos relacionales a Hadoop, se necesitaba una herramienta dedicada para realizar esta tarea rápidamente. Aquí es donde Apache Sqoop entró en escena, que ahora se usa ampliamente para transferir datos de archivos RDBMS al ecosistema de Hadoop para el procesamiento de MapReduce, etc.	• Podemos importar los resultados de una consulta SQL en HDFS usando Sqoop. • Para varios RDBMS, incluidos los servidores MySQL y Microsoft SQL, Sqoop ofrece conectores. • Sqoop es compatible con el protocolo de autenticación de red informática Kerberos, lo que permite que los nodos autentiquen a los usuarios mientras se comunican de forma segura a través de una red insegura. • Con un solo comando, Sqoop puede cargar la tabla completa o secciones	• La transferencia de datos es en paralelo, lo que la hace rápida y rentable. • Se pueden automatizar muchos procesos, lo que aumenta la eficiencia. • Es posible integrarse con la autenticación de seguridad de Kerberos . • Puede cargar datos directamente desde Hive y HBase. • Es una herramienta muy robusta con una gran comunidad de apoyo. • Se actualiza periódicamente, gracias a su continua contribución y desarrollo.	• Para realizar el análisis, ejecuta varios trabajos de reducción de mapas y, a veces, esto puede llevar mucho tiempo cuando hay muchas uniones si los datos no están normalizados. • Al usarse para la transferencia masiva de datos, podría ejercer una presión indebida sobre el almacén de datos de origen, y esto no es ideal si estos almacenes son muy utilizados por la aplicación comercial principal.	consigue aprovechar el entorno distribuido de nuestro cluster Hadoop obteniendo un rendimiento óptimo.

		específicas de la tabla.			
Apache Flume	Flume es un servicio distribuido, confiable y disponible para recopilar, agregar y mover de manera eficiente grandes cantidades de datos de registro. Tiene una arquitectura simple y flexible basada en flujos de datos de transmisión. Es robusto y tolerante a fallas con mecanismos de confiabilidad ajustables y muchos mecanismos de conmutación por error y recuperación. Utiliza un modelo de datos extensible simple que permite la aplicación analítica en línea. Apache Flume es una herramienta de software distribuida y open source. Se encarga de recopilar, agregar y mover datos desde diversas fuentes hasta	1. Código abierto: Apache Flume es un sistema distribuido de código abierto. Por lo tanto, está disponible de forma gratuita. 2. Flujo de datos: Apache Flume permite a sus usuarios crear flujos multisalto, de entrada y de salida. También permite el enrutamiento contextual, así como rutas de respaldo (conmutación por error) para los saltos fallidos. 3. Confiabilidad: En apache flume, las fuentes transfieren eventos a través del canal. La fuente del canal coloca eventos en el canal que luego son consumidos por el sumidero. El sumidero	Algunas de las principales ventajas de Apache Flume que hicieron que se eligiera esta tecnología se detallan aquí en viñetas: • Fuente abierta. • Hay disponible muy buena documentación, con muchos ejemplos y patrones de cómo se pueden aplicar. • Alto rendimiento con baja latencia. • Configuración declarativa. • Intrínsecamente distribuido. • Altamente confiable, disponible y escalable (horizontalmente). • Altamente extensible y personalizable. • Menos costos de instalación, operación y mantenimiento. • El aspecto de enrutamiento contextual tiene una	Algunas de las limitaciones de Apache Flume son: 1. Garantía de pedido débil: Apache Flume ofrece garantías más débiles que los otros sistemas, como colas de mensajes en el caso de que los datos se muevan más rápidamente y para permitir una tolerancia a fallas más económica. En el modo de confiabilidad de extremo a extremo de Apache Flume, los eventos de canal se entregan al menos una vez, pero sin garantías de pedido. 2. Duplicación: Apache Flume no garantiza que los mensajes que lleguen sean 100% únicos. En muchos escenarios, los mensajes duplicados pueden aparecer. 3. Baja escalabilidad: La escalabilidad de Flume a menudo es baja porque para cualquier empresa,	Las empresas de comercio electrónico utilizan Apache Flume para analizar el comportamiento de los clientes de una región en particular. Flume, por otra parte, se usa en entornos Hadoop y Big Data para ingestar y agregar grandes cantidades de datos hacia un almacenamiento centralizado.

	almacenamientos de datos.	transfiere el evento al siguiente agente o al repositorio de la terminal (como HDFS). 4. Recuperabilidad: Los eventos de canal se organizan en un canal de canal en cada agente de canal. Esto gestiona la recuperación de la falla. Además, Apache Flume admite un canal de archivos duradero. Los canales de archivos pueden estar respaldados por el sistema de archivos local. 5. Flujo constante: Apache Flume ofrece un flujo de datos constante entre las operaciones de lectura y escritura. Cuando la velocidad a la que llegan los datos supera la velocidad de escritura de datos en el destino, Apache Flume actúa como mediador	subsección dedicada en este capítulo. Pero para que tenga un aviso, este es un aspecto de Flume para observar la carga útil (transmisión de datos o evento) y construir un enrutamiento que sea adecuado.	dimensionar el hardware de un Apache Flume típico puede ser complicado y, en la mayoría de los casos, es prueba y error. Debido a esto, el aspecto de escalabilidad de Flume a menudo está bajo la lente. 4. Problema de confiabilidad: El rendimiento que puede manejar Apache Flume depende en gran medida del almacenamiento de respaldo del canal. Por lo tanto, si la tienda de respaldo no se elige sabiamente, puede haber problemas de escalabilidad y confiabilidad. 5. Topología compleja: Tiene una topología compleja y la reconfiguración es un desafío.	

		entre los productores de datos y los almacenes centralizados. Por lo tanto, ofrece un flujo constante de datos entre ellos. 6. Latencia: Apache Flume se adapta a un alto rendimiento con una latencia más baja. 7. Facilidad de uso: Con Flume, podemos ingerir la transmisión de datos de múltiples servidores web y almacenarlos en cualquiera de los almacenes centralizados como HBase, Hadoop HDFS, etc. 8. Entrega confiable de mensajes: Todas las transacciones en Apache Flume están basadas en canales. Para cada mensaje, hay dos transacciones: una para el remitente y otra			

		para el receptor. Esto asegura la entrega confiable de mensajes.			

Conclusion

En los últimos años, más empresas han llegado a comprender la importancia de los datos como la forma más realista de obtener inteligencia comercial. Por lo tanto, la ingestión de datos se ha vuelto más popular. Sin embargo, la extracción manual de datos comerciales de una gran cantidad de fuentes presenta muchos desafíos, con efectos negativos en el tiempo y las finanzas. Pero esto se ha convertido en cosa del pasado con la llegada de las herramientas de ingestión de datos. Las herramientas de ingesta de datos son herramientas de software que extraen automáticamente datos de una amplia gama de fuentes de datos y facilitan la transferencia de dichos flujos de datos a una única ubicación de almacenamiento.

El proceso de importar, transferir, cargar y procesar datos para su uso posterior o almacenamiento en una base de datos se denomina ingestión de datos y esto implica cargar datos de una variedad de fuentes, alterar y modificar archivos individuales y formatearlos para que quepan en un documento más grande. La ingesta de datos puede ser continua, asíncrona, en tiempo real o por lotes y la fuente y el destino también pueden tener un formato o protocolo diferente, lo que requerirá algún tipo de transformación o conversión. Las herramientas de ingesta de datos proporcionan un marco que permite a las empresas recopilar, importar, cargar, transferir, integrar y procesar datos de una amplia gama de fuentes de datos. Facilitan el proceso de extracción de datos al admitir varios protocolos de transporte de datos.

Además de recopilar, integrar y procesar datos, las herramientas de ingestión de datos ayudan a las empresas a modificar y formatear los datos con fines analíticos y de almacenamiento. Con estas herramientas, los usuarios pueden ingerir datos en lotes o transmitirlos en tiempo real. La ingestión de datos en tiempo real significa importar los datos tal como los produce la fuente. Por otro lado, ingerir datos en lotes significa importar fragmentos discretos de datos a intervalos.

Las empresas que utilizan herramientas de ingesta de datos deben priorizar las fuentes de datos, validar cada archivo y enviar elementos de datos al destino correcto para garantizar un proceso de ingesta eficaz. Aunque algunas empresas desarrollan sus propias herramientas, la mayoría utiliza herramientas de ingesta de datos desarrolladas por expertos en integración de datos. Las herramientas de ingesta de datos recopilan datos de varias fuentes y formatos en un repositorio central. Por ejemplo, reunir los datos de su software CRM y de servicio al cliente en un almacén de datos es un caso de uso de ingesta de datos.

La ingestión de datos es un paso crucial en la creación de cualquier plataforma de datos, ya que ayuda a eliminar los silos y compilar todos los datos de la organización en un solo lugar. Una vez que su herramienta de ingestión recopila todos los datos con éxito, puede comenzar a procesar y analizar esos datos para extraer información valiosa.

Referencias Bibliográficas

[1]"Data ingestion". www.cognizant.com. https://www.cognizant.com/us/en/glossary/data-ingestion#:~:text=Data%20ingestion%20is%20the%20process,can%20be%20accessed%20and%20analyzed. (accedido el 1 de marzo de 2023).

[2]"¿Qué es la ingesta de datos y por qué es importante esta tecnología". Striim. https://www.striim.com/blog/what-is-data-ingestion-and-why-this-technology-matters/ (accedido el 1 de marzo de 2023).

[3]simpleaprender "¿Qué es la ingesta de datos? Herramientas, tipos y conceptos clave | Simplilearn". Simplilearn.com. https://www.simplilearn.com/data-ingestion-article (accedido el 1 de marzo de 2023).

[4]"Principales herramientas de ingesta de datos en 2023". Aprender | Hevo. https://hevodata.com/learn/data-ingestion-tools/ (consultado el 1 de marzo de 2023).

[5]"Las mejores herramientas de ingesta de datos en 2023 | Una guía de comparación". atlán | Activa tus Metadatos. https://atlan.com/data-ingestion-tools/ (accedido el 1 de marzo de 2023).

[6]"Las 16 mejores herramientas de ingesta de datos para impulsar su estrategia de datos". Mejorado. https://improvado.io/blog/top-data-ingestion-tools (accedido el 1 de marzo de 2023).

[7]"Las 18 principales herramientas de ingesta de datos en 2022: revisiones, características, precios, comparación - INVESTIGACIÓN PAT: revisiones B2B, guías de compra y mejores prácticas". INVESTIGACIÓN PAT: Reseñas B2B, guías de compra y mejores prácticas. https://www.predictiveanalyticstoday.com/data-ingestion-tools/ (accedido el 1 de marzo de 2023).

[8]"Ingestión de datos: herramientas, tipos y conceptos clave | StreamSets". StreamSets. https://streamsets.com/learn/data-ingestion/ (accedido el 1 de marzo de 2023).

[9]"Apache Kafka". Apache Kafka. https://kafka.apache.org/intro (accedido el 1 de marzo de 2023).

[10]"Características de Kafka: aprendizaje simplificado". Aprendizaje simplificado. https://www.waytoeasylearn.com/learn/kafka-features/ (accedido el 1 de marzo de 2023).

[11]"4 Beneficios de Apache Kafka vs AMQP o JMS". Precisamente. https://www.precisely.com/blog/big-data/4-benefits-of-using-apache-kafka-in-lieu-of-amqp-or-jms (accedido el 1 de marzo de 2023).

[12]"Sqoop -". Sqoop -. https://sqoop.apache.org/ (accedido el 1 de marzo de 2023).

[13]"Qué es Sqoop - Introducción a Apache Sqoop - Intellipaat". Blog de Intellipaat. https://intellipaat.com/blog/what-is-apache-sqoop/ (accedido el 1 de marzo de 2023).

[14]"Lago de datos para empresas". Aprendizaje en línea de O'Reilly. https://www.oreilly.com/library/view/data-lake-for/9781787281349/38e2abc3-fba7-48bc-bbff-19cd74558a67.xhtml (accedido el 1 de marzo de 2023).

[15]"Bienvenido a Apache Flume — Apache Flume". Bienvenido a Apache Flume — Apache Flume. https://flume.apache.org/ (accedido el 1 de marzo de 2023).

[16]"Características y limitaciones de Apache Flume - DataFlair". Estilo de datos. https://data-flair.training/blogs/flume-features-limitations/ (accedido el 1 de marzo de 2023).

INSTITUTO TECNOLÓGICO SUPERIOR DE SAN ÁNDRES TUXTLA
TECNICAS DE ANÁLISIS, MINERIA Y VISUALIZACIÓN **INGENIERIA INFORMATICA**

PRIMER PARCIAL Fecha:28\02\2023

Nombre: Ricardo Mezo Bustamante Calificación:

I. Responde correctamente lo que a continuación se te pide (40%):

Describe las etapas o fases del Ciclo de vida de datos, considerando el objetivo único o específico de cada una de ellas.

Generación: La generación de datos ocurre independientemente de si las personas somos consciente de ello, especialmente en nuestro mundo cada vez más en línea. Algunos de estos datos son generados por las organizaciones, y algunos por terceros de los que puede o no tener conocimiento. Cada venta, compra, alquiler, comunicación, interacción, todo genera datos.

Captura: Cuando hablamos de captura, nos referimos al proceso de extraer información de documentos en papel o electrónicos y convertirla en datos para sistemas clave . Es donde la mayoría de las organizaciones comienzan su viaje de gestión de la información y transformación digital.

Almacenamiento: Esto se logra más comúnmente a través de la creación de bases de datos o conjuntos de datos. Estos conjuntos de datos se pueden almacenar en la nube, en servidores o mediante otra forma de almacenamiento físico, como un disco duro, un CD, un casete o un disquete. Al determinar cómo almacenar mejor los datos para las empresas, es importante incorporar un cierto nivel de redundancia para garantizar que una copia de sus datos esté protegida y accesible, incluso si la fuente original se corrompe o se ve comprometida.

Procesamiento: Una vez recopilados los datos, estos deben ser procesados. El procesamiento de datos puede referirse a varias actividades, por ejemplo, la gestión de datos, en la que un conjunto de datos se limpia y se transforma desde su forma original en algo más accesible y utilizable. Esto también se conoce como limpieza de datos, eliminación de datos o corrección de datos. También la compresión de datos , en la que los datos se transforman en un formato que se puede almacenar de manera más eficiente. Otra forma sería el cifrado de datos, en el que los datos se traducen a otra forma de código para protegerlos de problemas de privacidad.

Análisis: En esta parte nos referimos a los procesos que intentan obtener información significativa a partir de datos sin procesar. Los analistas y científicos de datos utilizan diferentes herramientas y estrategias para realizar estos análisis. Algunos de los métodos más utilizados incluyen modelos estadísticos, algoritmos, inteligencia artificial, minería de datos y aprendizaje automático.

Visualización: Una ves hemos realizado el análisis de datos y tenemos la información se busca una forma de presentarlos, esto se hace mediante la visualización, que es el proceso de creación de representaciones gráficas de su información, generalmente mediante el uso de una o más herramientas de visualización . La visualización de datos facilita la comunicación rápida de su análisis a un público más amplio, tanto dentro como fuera de su organización. La forma que toma su visualización depende de los datos con los que está trabajando, así como de la historia que desea comunicar.

Publicación: La publicación de datos es el proceso de hacer que la información, en particular los datos generados a partir de la investigación, estén disponibles para todos . El archivo de datos es el almacenamiento a largo plazo de tales datos y métodos. En ciencia, la publicación y el archivo de datos es importante para preservar la información científica para futuras investigaciones.

Implementación: En esta fase se brinda la oportunidad de dar sentido a su análisis y visualización. Más allá de simplemente presentar los datos, esto es cuando los investiga a través de la lente de su experiencia y comprensión. Es posible que su interpretación no solo incluya una descripción o explicación de lo que muestran los datos , sino, lo que es más importante, cuáles pueden ser las implicaciones.