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): **ATAXCA GOXCON FRANCISCO JAVIER**		Firma del alumno(s):
Producto: Investigación Unidad I	Nombre del Proyecto: Investigar (Tecnología de Generación y Captura de Datos)	Fecha: 12/ Febrero /2024
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%**		

Instituto Tecnológico Superior de San Andrés Tuxtla

Ingeniería Informática

Técnicas de análisis, minería y visualización

Octavo Semestre

Alumno: Francisco Javier Ataxca Goxcon

Docente: M.T.I. Lorenzo de Jesús Organista Oliveros

Introducción

En la presente investigación vamos a ver como la captura de datos es fundamental y un buen pilar para la buena gestión de los datos. La cantidad de información que continúa llegando a las empresas en formato papel o por correo electrónico es muy elevada. Por este motivo, la necesidad de disponer de sistemas automatizados de captura de datos resulta tan esencial. En la actualidad y dependiendo de la industria, este tipo de contenido puede llegar a representar entre el 80%-90% del total, por lo que si no se cuenta con las aplicaciones adecuadas para poder procesarlo se puede estar desperdiciando información de valor incalculable para la organización. En plena era *big data*, continuar anclados a los procesos manuales para introducir la información en nuestros sistemas nos penaliza en costes, rendimiento y eficiencia.

Objetivo.

Que el alumno tenga los conocimientos básicos teóricamente ante los temas que se abordan durante la unidad correspondiente y así tener un buen desempeño en las clases y poder entender de mejor manera al docente.

En una época en la que los volúmenes de datos aumentan rápidamente, es indispensable que las empresas recopilen y evalúen la información de forma eficiente. Solo así pueden obtener información valiosa sobre sus procesos empresariales y tomar decisiones con conocimiento de causa.

A medida que la IA sigue desarrollándose, la captura de datos desempeña un papel importante en este contexto. Las empresas pueden utilizar sistemas de captura de datos para recopilar información sobre sus clientes, productos, servicios y otros aspectos de su negocio e identificar procesos ineficientes.

La captura de datos puede definirse como el proceso de recopilación de datos de diversas fuentes y su conversión a un formato digital. Por ejemplo, estos datos pueden proceder de

- documentos como escaneado, foto, TIF o PDF nativo,
- Correos electrónicos,
- Formularios web,
- Plataformas de medios sociales
- y otras fuentes digitales

El proceso de captura de datos describe la recogida de datos en una empresa u organización. Consiste en recopilar información de distintas fuentes y almacenarla en un formato uniforme. Esto ocurre en 4 pasos:

1. Establecer datos

   En primer lugar, las empresas tienen que decidir qué datos quieren capturar en línea en un sistema con Captura de Datos. Aquí es importante recopilar solo la información relevante para que el proceso sea lo más eficiente posible.

2. Identificar las fuentes de datos

   A continuación, las empresas deben identificar las fuentes de datos de las que quieren extraer la información. Puede tratarse de fuentes internas, como bases de datos, o externas, como sitios web.

3. Introducir datos

   Una vez identificada toda la información pertinente, las empresas la capturan. Pueden hacerlo manualmente o mediante la automatización de la captura de datos. Con la captura manual, los datos deben ser introducidos en el sistema por un empleado. Con la captura automatizada, se utiliza un software de captura de datos que extrae automáticamente los datos de las fuentes.

4. Guardar datos

   Una vez recopilados los datos, las empresas deben almacenarlos en un formato coherente. Aquí es importante que el formato sea uniforme para todas las fuentes de datos, de modo que éstos puedan analizarse y procesarse fácilmente más adelante. Aquí es donde la gestión de la captura de datos desempeña un papel crucial. Garantiza que todos los datos se almacenan de forma correcta, uniforme y adecuada.

Rápidos avances tecnológicos

Si echamos la vista atrás, los inicios en reconocimiento óptico de caracteres (OCR) pueden encontrarse en el siglo XIX, con los primeros escáneres de retina de Charles Carey en 1870. Podría situarse en esa fecha el punto de partida de una tecnología cuyos avances han sido extraordinarios, siendo ya en la década de 1970 cuando la tecnología en el reconocimiento de caracteres había evolucionado tanto que era capaz de leer texto manuscrito.

Los sistemas OCR más avanzados a finales del siglo XX se basaban en plantillas para ofrecer resultados consistentes en el reconocimiento de caracteres. Entre las desventajas de este modelo destacaba la cantidad de configuración previa que era necesario realizar, indicando, entre otras, las áreas exactas en las que se encontraba el texto, de manera que a cada cambio en el diseño de los textos era necesario reconfigurar las plantillas, para lo que se requería de personal cualificado.

La captura inteligente

La incorporación de la tecnología de Inteligencia Artificial (IA) y aprendizaje automático (Machine Learning, ML) a los sistemas OCR ha sido lo que ha marcado un antes y un después en lo que a captura de datos se refiere. Pasamos de un mero reconocimiento óptico de caracteres a una tecnología de captura inteligente de datos que, obviamente, es un campo mucho más amplio de recopilación y análisis de información.

Si un sistema OCR era capaz de introducir en el sistema resultados como fechas con formato “02-03-2021” con una precisión del 100%, la tecnología de captura inteligente proporciona significado al texto extraído a partir de todo tipo de contenido no estructurado, contextualizando fechas de vencimiento de pago, fechas de entrada o salida de pedidos, etc.

Los nuevos sistemas de captura de datos leen y comprenden los documentos digitales prácticamente del mismo modo que los seres humanos y, como ellos, tienen 'capacidad de aprender'. En la actualidad, podemos entrenar algoritmos inteligentes para buscar entidades específicas como fechas, números de contrato o números de órdenes de compra en diferentes documentos. Estos sistemas entrenados producen regularmente niveles de precisión de más del 90%. Por lo tanto, uno puede analizar de manera eficiente cientos y miles de documentos por minuto.

Tecnologías populares para la captura de datos.

Introducción manual de datos

En esta forma de captura de datos, las empresas introducen manualmente los datos -por ejemplo, de formularios- en un ordenador para digitalizarlos. Sin embargo, este método de captura de datos sólo es adecuado para una empresa que necesite capturar y procesar un volumen bajo y variable de datos. Esto se debe a que la captura manual de datos depende del trabajo humano y, por tanto, es propensa a errores**.**

OCR - Reconocimiento óptico de caracteres

OCR es un ejemplo sencillo de captura de datos para capturar textos completos. Se trata de una tecnología que reconoce caracteres y fuentes generados por máquinas. Las empresas pueden utilizar el OCR para extraer y procesar automáticamente texto de documentos escaneados y archivos PDF, por ejemplo. El OCR se utiliza a menudo cuando se generan grandes volúmenes de datos similares, por ejemplo, en los sectores sanitario, asegurador y financiero. El OCR suele complementarse con soluciones ICR, IDP u OMR.

ICR - Reconocimiento inteligente de caracteres

ICR puede leer caracteres manuscritos de cualquier fuente y convertirlos en datos significativos. Por ejemplo, ICR prepara datos manuscritos de formularios para que una empresa pueda procesarlos fácilmente. Esta tecnología la utilizan sobre todo bancos y entidades financieras. ICR es la nueva generación de tecnología OCR.

Códigos de barras y códigos QR

La tecnología del código de barras puede leer la información de los códigos de barras y convertirla en formato digital. Hay que distinguir entre códigos de barras 1D y códigos de barras 2D. Los códigos de barras 1D se utilizan en tiendas, por ejemplo, para hacer un seguimiento del inventario. También se utilizan en hospitales para comprobar los datos de los pacientes. Los códigos de barras 2D -también llamados códigos de respuesta rápida- son adecuados para capturar páginas web o documentos, por ejemplo. En la práctica, es el caso de la publicidad y los envases de productos, por ejemplo.

Conclusión

Para concluir La captura de datos es un proceso importante para que las empresas recopilen y analicen información valiosa y extraigan las conclusiones adecuadas para sus operaciones empresariales diarias y, de este modo, logren mejores resultados empresariales.

Fuentes:

https://konfuzio.com/es/captura-de-datos/#data-capture-definition

https://www.adea.es/blog/captura-de-datos/

https://www.tecnologias-informacion.com/captura-datos.html

GUIA DE OBSERVACIÓN PARA RESOLUCIÒN DE EJERCICIOS PRACTICOS

ITSSAT

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

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

ALUMNO: **ATAXCA GOXCON FRANCISCO JAVIER**

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%**		

Instituto Tecnológico Superior de San Andrés Tuxtla

Ingeniería Informática

Técnicas de análisis, minería y visualización

Octavo Semestre

Alumno: Francisco Javier Ataxca Goxcon

Docente: M.T.I. Lorenzo de Jesús Organista Oliveros

Herramientas de ingesta

La ingestión de datos se define como el proceso de absorber datos de una variedad de fuentes y transferirlos a un sitio de destino donde pueden depositarse y analizarse. En términos generales, los destinos pueden ser una base de datos, un almacén de datos, un almacén de documentos, un mercado de datos, etc. Por otro lado, existen varias opciones de origen, como hojas de cálculo, extracción o scrapping de datos web, aplicaciones internas y SaaS. datos.

Los datos empresariales generalmente se almacenan en múltiples fuentes y formatos. Por ejemplo, los datos de ventas se almacenan en Salesforce.com, los DBMS relacionales almacenan información del producto, etc. Como estos datos se originan en diferentes ubicaciones, deben limpiarse y convertido en una forma que se pueda analizar fácilmente para la toma de decisiones utilizando una herramienta de ingestión de datos fácil de usar. De lo contrario, se quedará con piezas de rompecabezas que no se pueden unir.

La ingesta de datos se puede realizar de diferentes maneras, como en tiempo real, por lotes o una combinación de ambos (conocida como arquitectura lambda) según los requisitos comerciales. Veamos formas de realizarlo con más detalle.

Las herramientas de carga tradicionales de datos suelen ser adecuadas para cargas de datos relativamente pequeñas y estructuradas, donde la información se extrae de bases de datos relacionales y se carga en un sistema de destino. Por otro lado, las herramientas de ingesta en Big Data son herramientas especializadas que se utilizan para cargar grandes volúmenes de datos no estructurados y semiestructurados en sistemas de Big Data en tiempo real.

Aspecto	Kafka	Sqoop	Flume
Tipo de operación de datos	Kafka se utiliza para construir pipelines de datos en tiempo real que transfieren datos entre sistemas o aplicaciones, transforman flujos de datos o reaccionan ante ellos. Funciona como un sistema de mensajería para publicar y suscribirse a flujos de registros.	Sqoop se utiliza para la transferencia masiva de datos entre Hadoop y bases de datos relacionales. Admite tanto la importación como la exportación de datos.	Flume se utiliza para recopilar y transferir grandes cantidades de datos a un almacén de datos centralizado. Aunque está diseñado principalmente para datos de registro, también puede utilizarse para otros tipos de fuentes de datos, como datos de eventos, tráfico de red e incluso mensajes de correo electrónico.
Herramientas y componentes	Kafka se basa en temas, productores y consumidores.	Sqoop proporciona herramientas de importación y exportación para mover datos entre bases de datos y HDFS.	Flume utiliza una combinación de fuente-canal-sumidero configurada como un agente Flume en un archivo de configuración.
Estado de ejecución	Los procesos de importación/exportación de Sqoop terminan después de la transferencia de datos.	Un agente Flume transmite los datos disponibles cuando se inicia y continúa ejecutándose para transmitir nuevos datos a medida que están disponibles.	Un productor/consumidor de Kafka también continúa ejecutándose y transmitiendo mensajes en tiempo real a medida que se publican en un tema.
Soporte para múltiples agentes o suscriptores	Sqoop no admite la vinculación de procesos de importación o exportación para formar un proceso de importación/exportación múltiple.	Flume admite un agente de múltiples flujos en el que la salida de un agente puede utilizarse como entrada para otro agente.	Kafka es una plataforma de múltiples suscripciones, lo que significa que varios productores y consumidores pueden suscribirse al mismo tema simultáneamente

Flexibilidad	Adaptable a diversos formatos de datos, permite el procesamiento en tiempo real y la persistencia.	Enfocado en transferencia entre bases de datos relacionales y Hadoop.	Especializado en logs y eventos, puede ser personalizado para diferentes fuentes y destinos.
Escalabilidad	Muy alta, puede manejar grandes volúmenes de datos y transacciones en tiempo real.	Alta, puede manejar grandes volúmenes de datos en operaciones de carga masiva.	Moderada, es capaz de manejar flujos de datos, pero puede tener limitaciones en grandes volúmenes

Fuentes:

[1] Diego Calvo. Jul 5 208. DiegoCalvo. Herramientas de ingesta. Disponible: https://www.diegocalvo.es/big-data-herramientas-de-ingesta-de-datos/

[2] Corporativo. 6/2017. Blog. Disponible: https://bigdatadummy.wordpress.com/2017/11/06/kafka-vs-flume-vs-spark/

[3] Rubén Gerardo. 4/05/2023. Linkedin. Herramientas de carga de información. Disponible: https://es.linkedin.com/pulse/herramientas-de-carga-informac%C3%ADon-rub%C3%A9n-gerardo-carmona-pozos#:~:text=Por%20otro%20lado%2C%20las%20herramientas,Big%20Data%20en%20tiempo%20real.

[4] Corporativo. 12/01/2023. Astera. Ingestión de datos. Disponible: https://www.astera.com/es/type/blog/data-ingestion/

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

PRIMER PARCIAL Fecha:14\02\2024
Nombre: Francisco Javier Ataxca Goxcon **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.

1.-Generación.

Esta es la primera etapa del ciclo de vida de los datos en la cual normalmente incluye la intervención humana.
En esta fase, lo más importante es establecer el formato de los datos que constituyen la prueba del dato que acaba de generar, incluidos aspectos como su codificación, la precisión, las unidades, el formato, etc.

2.- Captura.

En esta fase, nos centraremos en herramientas que permiten acceder a datos más o menos estructurados, los cuales pueden existir durante un periodo limitado y que hay que ir capturando conforme van apareciendo. Por lo tanto, el objetivo de esta fase es capturar los datos de una forma estructurada y sea entendible para las personas que ocuparan la información.

3- Almacenamiento.

En esta etapa del ciclo de vida de los datos es claro ya que el objetivo principal es adoptar soluciones de alanceamiento para poder tener un respaldo y así poder tener registros históricos con los que se pueden hacer futuros análisis con datos pasados.

4.- Procesamiento.

Normalmente es necesario realizar un trabajo previo de procesado de los datos capturados y/o almacenados en bruto, para posteriormente poder analizar la información obtenida y ponerlos en valor.

En esta fase se incluyen todas las herramientas y procesos orientados a la limpieza de datos, su selección, la fusión de diferentes fuentes de datos o su

enriquecimiento mediante el uso de datos de terceros o API, etc., con el objetivo de disponer de un único fichero para su análisis posterior.

5.- Análisis.

Esta fase es la que se considera como clave en cualquier proyecto de ciencia de datos. Durante esta fase, es posible que tengamos que realizar diferentes tipos de análisis estadísticos (descriptivos, inferenciales, etc.) o de minería de datos (supervisados o no supervisados), así como un análisis visual para explorar e inspeccionar los datos en función de nuestro objetivo, aprendiendo de la naturaleza de los datos con los que estemos trabajando.

El objetivo final será poner en valor los datos y poder obtener conclusiones sobre ellos mediante la construcción de modelos que nos permitan predecir valores futuros, la generación de clasificaciones o agrupaciones, etc.

6.- Visualización.

La fase de la visualización nos permite también la presentación de los datos y resultados obtenidos de forma gráfica, de manera que sean fácilmente interpretables tanto para los científicos de datos como para el usuario final. Actualmente, las herramientas y recursos que tenemos a nuestro alcance nos facilitan poder difundir las conclusiones de la fase de análisis, cosa que permite, incluso, poder navegar o profundizar en los datos presentados mediante la interacción con ellos.

7.- Publicación.

En esta fase, el objetivo principal es poner a disposición de terceros las conclusiones o resultados en forma de nuevos datos.
Actualmente, existen diferentes formas de llevarlo a cabo, como la publicación de tu propia web o webs de terceros, compartir la información en sistemas de almacenamiento en la nube, etc., pero lo más recomendable es utilizar un repositorio digital pensado para dicho uso.

8.- Implementación

Finalmente, aunque no se trata de una etapa del ciclo de vida propiamente dicha, es necesario poder convertir un modelo creado en un «laboratorio» en una herramienta o un producto que resuelva el problema para el cual fue creado,