1.1. Recopilación, integración y almacenamiento de datos

May 20, 2025

Creado por:

Isabel Maniega

1 1.1 Recopilación, integración y almacenamiento de datos

1.1 1.1.1 ¿Qué es la recopilación de datos?

La recopilación de datos es el proceso sistemático de recopilación y registro de información o datos de diversas fuentes para su análisis, interpretación y toma de decisiones. Es un paso fundamental en la investigación, las operaciones comerciales y prácticamente todos los campos en los que se utiliza la información para comprender, mejorar o tomar decisiones informadas.

1.2 Elementos clave de la recopilación de datos

- **Fuentes:** Los datos se pueden recopilar de una amplia gama de fuentes, incluidas encuestas, entrevistas, observaciones, sensores, bases de datos, redes sociales y más.
- **Métodos:** Se emplean varios métodos para recopilar datos, como cuestionarios, ingreso de datos, extracción de datos web y redes de sensores. La elección del método depende del tipo de datos, los objetivos de la investigación y los recursos disponibles.
- **Tipos de datos:** Los datos pueden ser cualitativos (descriptivos) o cuantitativos (numéricos), estructurados (organizados en un formato predefinido) o no estructurados (texto o medios de formato libre), y primarios (recopilados directamente) o secundarios (obtenidos de fuentes existentes).
- **Herramientas de recopilación de datos:** La tecnología desempeña un papel importante en la recopilación de datos moderna, con aplicaciones de software, aplicaciones móviles, sensores y plataformas de recopilación de datos que facilitan la captura eficiente y precisa de datos.
- **Consideraciones éticas:** Se deben seguir las pautas éticas, incluido el consentimiento informado y la protección de la privacidad, para garantizar que la recopilación de datos respete los derechos y el bienestar de las personas.
- **Calidad de los datos:** La precisión, integridad y confiabilidad de los datos recopilados son fundamentales para su utilidad. Se implementan medidas de garantía de calidad de los datos para minimizar los errores y sesgos.
- **Almacenamiento de datos:** Los datos recopilados deben almacenarse y administrarse de forma segura para evitar pérdidas, acceso no autorizado y violaciones. Las soluciones de almacenamiento de datos varían desde servidores locales hasta plataformas basadas en la nube.

1.3 Importancia de la recopilación de datos en las empresas modernas

La recopilación de datos es de suma importancia en las empresas modernas por varias razones convincentes:

- **Toma de decisiones informada:** Los datos recopilados sirven como base para la toma de decisiones informada en todos los niveles de una organización. Proporcionan información valiosa sobre el comportamiento del cliente, las tendencias del mercado, la eficiencia operativa y más.
- **Ventaja competitiva:** Las empresas que recopilan y analizan datos de manera eficaz obtienen una ventaja competitiva. La información basada en datos ayuda a identificar oportunidades, optimizar procesos y mantenerse por delante de los competidores.
- **Comprensión del cliente:** La recopilación de datos permite a las empresas comprender mejor a sus clientes, sus preferencias y sus puntos débiles. Esta información es invaluable para adaptar productos, servicios y estrategias de marketing.
- **Medición del rendimiento:** La recopilación de datos permite a las organizaciones evaluar el rendimiento de varios aspectos de sus operaciones, desde campañas de marketing hasta procesos de producción. Esto ayuda a identificar áreas de mejora.
- **Gestión de riesgos:** Las empresas pueden utilizar los datos para identificar riesgos potenciales y desarrollar estrategias para mitigarlos. Esto incluye riesgos financieros, interrupciones en la cadena de suministro y amenazas de ciberseguridad.
- **Innovación:** La recopilación de datos respalda la innovación al brindar información sobre tendencias emergentes y demandas de los clientes. Las empresas pueden usar esta información para desarrollar nuevos productos o servicios.
- **Asignación de recursos:** La toma de decisiones basada en datos ayuda a asignar recursos de manera eficiente. Por ejemplo, los presupuestos de marketing se pueden optimizar en función del rendimiento de diferentes canales.

1.4 Metas y objetivos de la recopilación de datos

Las metas y objetivos de la recopilación de datos dependen del contexto específico y de las necesidades de la organización o del proyecto de investigación. Sin embargo, existen algunos objetivos generales comunes:

- **Recopilación de información:** el objetivo principal es recopilar información precisa, relevante y confiable que aborde preguntas u objetivos específicos.
- **Análisis y conocimiento:** los datos recopilados están destinados a ser analizados para descubrir patrones, tendencias, relaciones y conocimientos que puedan informar la toma de decisiones y el desarrollo de estrategias.
- **Medición y evaluación:** la recopilación de datos permite la medición y evaluación de varios factores, como el rendimiento, la satisfacción del cliente o el potencial del mercado.
- **Resolución de problemas:** la recopilación de datos puede estar dirigida a resolver problemas o desafíos específicos que enfrenta una organización, como identificar las causas fundamentales de los problemas de calidad.
- **Monitoreo y vigilancia:** en algunos casos, la recopilación de datos sirve como una función de monitoreo o vigilancia continua, que permite a las organizaciones realizar un seguimiento de los procesos o las condiciones en curso.
- **Evaluación comparativa:** La recopilación de datos se puede utilizar para realizar evaluaciones comparativas con los estándares de la industria o con la competencia, lo que ayuda a

las organizaciones a evaluar su desempeño en relación con otros.
- **Planificación y estrategia:** Los datos recopilados a lo largo del tiempo pueden respaldar la planificación a largo plazo y el desarrollo de estrategias, lo que garantiza que las organizaciones se adapten a las circunstancias cambiantes.

En resumen, la recopilación de datos es una actividad fundamental con diversas aplicaciones en diferentes industrias y sectores. Sus objetivos van desde comprender a los clientes y tomar decisiones informadas hasta mejorar los procesos, gestionar los riesgos e impulsar la innovación. La calidad y la relevancia de los datos recopilados son fundamentales para lograr estos objetivos.

1.5 ¿Cómo planificar su estrategia de recopilación de datos?

Antes de comenzar, revisaremos los pasos cruciales de la planificación de su estrategia de recopilación de datos. Su éxito en la recopilación de datos depende en gran medida de qué tan bien defina sus objetivos, seleccione las fuentes adecuadas, establezca metas claras y elija los métodos de recopilación apropiados.

1.5.1 Definición de las preguntas de investigación

Definir las preguntas de investigación es la base de cualquier esfuerzo eficaz de recopilación de datos. Cuanto más precisas y relevantes sean las preguntas, más valiosos serán los datos que recopile.

- **La especificidad es clave:** Asegúrese de que las preguntas de investigación sean específicas y estén centradas. En lugar de preguntar "¿Cómo podemos mejorar la satisfacción del cliente?", pregunte "¿Qué aspectos específicos de nuestro servicio encuentran los clientes más satisfactorios o insatisfactorios?".
- **Priorizar las preguntas:** Determine las preguntas más críticas que tendrán el mayor impacto en sus objetivos. No todas las preguntas son igualmente importantes, por lo que debe asignar sus recursos en consecuencia.
- **Alineación con los objetivos:** Asegúrese de que las preguntas de investigación se alineen directamente con sus objetivos generales. Si su objetivo es aumentar las ventas, las preguntas de investigación deben estar orientadas a comprender los comportamientos y las preferencias de compra de los clientes.

1.5.2 Identificación de fuentes de datos clave

Identificar las fuentes de datos adecuadas es esencial para recopilar información precisa y relevante. A continuación, se muestran algunos ejemplos de fuentes de datos clave para diferentes industrias y propósitos.

- **Datos de clientes:** Esto puede incluir datos demográficos de los clientes, historial de compras, comportamiento del sitio web y comentarios de las interacciones de servicio al cliente.
- **Informes de investigación de mercado:** Utilice informes de la industria, análisis de la competencia y estudios de tendencias del mercado para recopilar datos e información externos.
- **Registros internos:** Las bases de datos, los registros financieros y los datos operativos de su organización pueden brindar información valiosa sobre el desempeño de su negocio.
- **Plataformas de redes sociales:** Monitoree los canales de redes sociales para recopilar comentarios de los clientes, realizar un seguimiento de las menciones de la marca e identificar tendencias emergentes en su industria.

- **Análisis web:** Recopile datos sobre el tráfico del sitio web, el comportamiento del usuario y las tasas de conversión para optimizar su presencia en línea.

1.5.3 Establecer objetivos claros de recopilación de datos

Establecer objetivos claros y mensurables es esencial para garantizar que sus esfuerzos de recopilación de datos sigan por el buen camino y brinden resultados valiosos. Los objetivos deben ser:

- **Específicos:** Defina claramente lo que pretende lograr con la recopilación de datos. Por ejemplo, aumentar el tráfico del sitio web en un 20 % en seis meses es un objetivo específico.
- **Medibles:** Establezca criterios para medir su progreso y éxito. Utilice métricas como el crecimiento de los ingresos, las puntuaciones de satisfacción del cliente o las tasas de conversión.
- **Alcanzables:** Establezca objetivos realistas que su equipo pueda alcanzar de manera realista. Los objetivos demasiado ambiciosos pueden generar frustración y agotamiento.
- **Relevantes:** Asegúrese de que sus objetivos se alineen con los objetivos más amplios y las iniciativas estratégicas de su organización.
- **Limitar el tiempo:** Establezca un plazo dentro del cual planea alcanzar sus objetivos. Esto agrega una sensación de urgencia y lo ayuda a realizar un seguimiento del progreso de manera eficaz.

1.5.4 Elección de métodos de recopilación de datos

La selección de los métodos de recopilación de datos correctos es crucial para obtener datos precisos y confiables. Su elección debe estar alineada con sus preguntas y objetivos de investigación. A continuación, se ofrece una mirada más detallada a los distintos métodos de recopilación de datos y sus aplicaciones prácticas.

Antes de profundizar en los métodos de recopilación de datos, debe comprender las principales categorías de datos en sí. Hay dos categorías principales de datos: **Datos primarios** y **Datos secundarios**.

Los **Datos primarios** son el tipo que usted mismo recopila. Significa que participa activamente en la obtención de información. Puede proporcionar una pregunta de investigación y realizar experimentos por su cuenta. Su experimento puede proporcionarle resultados que pueden ser datos cualitativos o cuantitativos.

- **Datos cualitativos:** este tipo de datos no son cuantificables y no se presentan en números. Son datos que pueden informar la calidad de su producto, empresa o servicio con atributos cualitativos y explicables. Este tipo de datos es increíblemente útil y también difícil de recopilar y observar de manera eficiente (por eso el web scraping es esencial).

- **Datos cuantitativos**: como sugiere el título, se trata de la cantidad. Este tipo de datos implica cifras reales para enfatizar la cantidad de datos involucrados. Algunos ejemplos básicos de datos cuantitativos podrían ser la cantidad de personas que visitan un centro comercial los domingos, la cantidad de personas que migran a un área geográfica por año o la cantidad de trabajadores en una oficina.

Los **datos secundarios** están a nuestro alrededor. Son fácilmente accesibles en Internet y requieren menos recursos para recopilarlos, a diferencia de los datos primarios. En este caso, la recopilación de datos primarios la realizó otra persona antes de subirlos a Internet. Los datos secundarios vienen

en forma de resultados de búsqueda que brindan información sobre menciones de marca, análisis de precios, reseñas y sentimientos.

Entonces, ¿cómo puede obtener la categoría de datos que sea relevante para sus necesidades actuales? Aquí es donde entran en juego las herramientas de recopilación de datos. Eligir métodos y soluciones desarrollados para ayudar a recopilar datos de manera ordenada y que puedan analizarse, comprenderse e integrarse fácilmente en procesos, negocios o de otro tipo.

1.6 Consideraciones éticas en la recopilación de datos

Las consideraciones éticas son fundamentales en la recopilación de datos para garantizar la privacidad, la equidad y la transparencia. Abordar estas cuestiones no solo es responsable, sino también crucial para generar confianza con las partes interesadas.

1.6.1 Consentimiento informado

Obtener el consentimiento informado de los participantes es un imperativo ético. La transparencia es fundamental y los participantes deben comprender plenamente el propósito de la recopilación de datos, cómo se utilizarán sus datos y los posibles riesgos o beneficios involucrados. El consentimiento debe ser voluntario y los participantes deben tener la opción de retirar su consentimiento en cualquier momento sin consecuencias.

Los formularios de consentimiento deben ser claros y comprensibles, evitando un lenguaje excesivamente complejo o jerga legal. Se debe tener especial cuidado al recopilar datos sensibles o personales para garantizar que se respeten los derechos de privacidad.

1.6.2 Protección de la privacidad

Proteger la privacidad de las personas es esencial para mantener la confianza y cumplir con las regulaciones de protección de datos. Se debe utilizar la anonimización o seudonimización de datos para evitar la identificación de las personas, especialmente al compartir o publicar datos. Se deben implementar métodos de cifrado de datos para proteger los datos tanto en tránsito como en reposo, y resguardarlos del acceso no autorizado.

Se deben implementar controles de acceso estrictos para restringir el acceso a los datos solo al personal autorizado, y se deben establecer y respetar políticas claras de retención de datos, evitando el almacenamiento innecesario de datos. Se deben realizar auditorías de privacidad periódicas para identificar y abordar posibles vulnerabilidades o problemas de cumplimiento.

1.6.3 Sesgo y equidad en la recopilación de datos

Abordar el sesgo y garantizar la equidad en la recopilación de datos es fundamental para evitar la perpetuación de las desigualdades. Los métodos de recopilación de datos deben diseñarse para minimizar los sesgos potenciales, como el sesgo de selección o el sesgo de respuesta. Se deben realizar esfuerzos para lograr muestras diversas y representativas, asegurando que los datos reflejen con precisión la población de interés. Es esencial brindar un trato justo a todos los participantes y fuentes de datos, y evitar estrictamente la discriminación basada en características como la raza, el género o el nivel socioeconómico.

Si se utilizan algoritmos en la recopilación o el análisis de datos, se deben evaluar y mitigar los sesgos que puedan surgir de los procesos automatizados. Se pueden considerar revisiones éticas

o consultas con expertos cuando se trabaja con datos sensibles o potencialmente sesgados. Al adherirse a principios éticos durante todo el proceso de recopilación de datos, se protegen los derechos de las personas y se establece una base para la toma de decisiones responsable y confiable basada en datos.

1.6.4 Sesgo y representación

Garantizar que los métodos de recopilación de datos estén libres de sesgos y representen con precisión a poblaciones diversas es un desafío. Por ejemplo, las tecnologías de reconocimiento facial han enfrentado críticas por sesgo racial, donde ciertos grupos demográficos no son reconocidos con precisión, lo que genera preocupaciones éticas sobre la justicia y la igualdad.

1.6.5 Transparencia y rendición de cuentas

Mantener la transparencia en la forma en que se recopilan, utilizan y comparten los datos es un desafío, pero esencial para el cumplimiento ético. El desafío radica en comunicar prácticas de datos complejas de una manera comprensible para los usuarios. La falta de transparencia puede conducir a situaciones como el caso de Google Street View, donde Google fue criticado por recopilar más datos de los que reveló, incluidos detalles de la red Wi-Fi personal.

1.6.6 Cumplimiento legal y normativo

Navegar por el complejo panorama de las leyes internacionales de protección de datos, como el RGPD en Europa y las diferentes leyes en los distintos países, es un desafío importante para las organizaciones globales. El cumplimiento requiere una vigilancia constante y la adaptación a las normas legales en evolución.

1.6.7 Resumen: Ética de datos

Principios básicos de la ética de datos

- **Privacidad:** Proteger el derecho de las personas a controlar su información personal y garantizar que la recopilación de datos sea transparente y consensuada.
- **Seguridad:** Implementar medidas sólidas para proteger los datos de accesos no autorizados, infracciones y robos.
- **Justicia:** Garantizar que los datos se utilicen de manera justa y que no discriminen a ninguna persona o grupo.
- **Transparencia:** Hacer que los procesos de recopilación, análisis y uso de datos sean abiertos y comprensibles para todas las partes interesadas.
- **Responsabilidad:** Hacer que las organizaciones y las personas sean responsables de cómo recopilan, usan y comparten los datos.

Desafíos de la ética de datos

- **Equilibrar la innovación con la privacidad:** Encontrar el equilibrio adecuado entre aprovechar los datos para la innovación y respetar la privacidad de las personas puede ser un desafío.
- **Sesgo y discriminación:** Los datos y algoritmos pueden perpetuar sesgos inadvertidamente, lo que lleva a la discriminación si no se gestionan con cuidado.

- **Propiedad y control de los datos:** A medida que los datos se convierten en un activo valioso, determinar quién es su propietario y quién puede controlar su uso es cada vez más polémico.
- **Cumplimiento normativo:** Navegar por el complejo panorama de leyes y regulaciones de protección de datos en diferentes jurisdicciones se suma al desafío.

1.7 ¿Por qué es importante la recopilación precisa de datos?

La recopilación precisa de datos es crucial para garantizar la validez y confiabilidad de los hallazgos de la investigación. Apoya la toma de decisiones informada, mejora la precisión y exactitud, asegura la calidad y mantiene la integridad de la investigación. Sin datos precisos, las conclusiones extraídas de la investigación pueden ser erróneas o engañosas.

- **Toma de decisiones informada:** Los datos precisos proporcionan una base sólida para tomar decisiones que tienen más probabilidades de ser efectivas y beneficiosas. Ayudan a las partes interesadas a comprender el contexto y las implicaciones de sus elecciones.
- **Precisión y exactitud:** Los datos de alta calidad garantizan que los resultados de la investigación sean precisos y exactos, lo que genera confianza en los hallazgos y respalda la credibilidad de la investigación.
- **Garantía de calidad:** Garantizar la exactitud de los datos es esencial para mantener la calidad de la investigación. Ayuda a identificar y corregir errores, lo que conduce a resultados más confiables y válidos.
- **Integridad de la investigación** La recopilación precisa de datos mantiene la integridad del proceso de investigación. Garantiza que los hallazgos sean confiables y puedan ser replicados por otros investigadores.

1.8 El derecho a permanecer anónimo

Gracias a la legislación sobre privacidad de datos, encabezada por el RGPD de Europa y la CPRA de California, el consumidor ha obtenido voz y, con ella, el derecho a permanecer anónimo. De modo que, cuando una organización utilice mis datos (como inevitablemente hará), nunca podrán rastrearme hasta mí. Esta es la esencia de la anonimización de datos. La anonimización de datos es una categoría general que incluye el enmascaramiento de datos, la seudonimización, la agregación de datos, la aleatorización de datos, la generalización de datos y el intercambio de datos. Esta guía profundiza en cada una de estas técnicas de anonimización de datos y, a continuación, analiza los pros y los contras del proceso de anonimización, los desafíos que enfrenta y las direcciones para futuras investigaciones. Concluye revelando un enfoque innovador para garantizar la privacidad personal, el cumplimiento de las regulaciones, la confianza del cliente y el derecho a permanecer anónimo.

1.8.1 ¿Qué es la anonimización de datos?

La anonimización de datos es el proceso de ocultar o eliminar información de identificación personal (PII) de un conjunto de datos para proteger la privacidad de las personas asociadas con esos datos. La anonimización de los datos hace que sea imposible reconocer a las personas a partir de sus datos, al tiempo que mantiene la información funcional para pruebas de software, análisis de datos u otros fines legítimos. La anonimización de datos transforma la PII y los datos confidenciales de tal manera que no se puedan vincular fácilmente a una persona específica. En otras palabras, reduce el riesgo de reidentificación, con el fin de cumplir con las leyes de privacidad de datos y

aumentar la seguridad. El proceso de anonimización generalmente implica el enmascaramiento de datos PII, como nombres, direcciones, números de teléfono, detalles del pasaporte o números de la seguridad social. Con este fin, los valores se reemplazan o eliminan, mediante el uso de técnicas criptográficas, o agregando ruido aleatorio, para proteger los datos. Los datos anonimizados no pueden garantizar un anonimato completo, con la amenaza de reidentificación, en particular cuando los datos anonimizados se combinan con fuentes disponibles públicamente. Por lo tanto, los equipos de datos deben considerar cuidadosamente los riesgos y las limitaciones de sus herramientas y procesos de anonimización de datos cuando trabajan con datos personales o confidenciales.

1.8.2 El papel que desempeña la anonimización de datos en la protección de la privacidad personal

La anonimización de datos desempeña un papel fundamental en la protección de la privacidad personal al evitar la exposición y la explotación de la información confidencial de las personas. Con la cantidad cada vez mayor de datos que se recopilan y almacenan, el riesgo de que se pueda acceder a la información personal y hacer un uso indebido de ella (sin el conocimiento o el consentimiento de alguien) es mayor que nunca. Cuando se viola la información personal, no solo se trata de una violación de la seguridad para la organización, sino, lo que es más importante, de una violación de la confianza para el cliente o consumidor. Estos ataques pueden dar lugar a amplias violaciones de la privacidad, como incumplimiento de contrato, discriminación y robo de identidad. Al ocultar o eliminar la información de identificación personal de los conjuntos de datos, la anonimización de datos limita gravemente la capacidad de los usuarios no autorizados de acceder o utilizar la información personal. Además de prevenir violaciones de la privacidad y proteger los derechos de las personas, la anonimización de datos permite a las organizaciones cumplir con las regulaciones de privacidad de datos (como APPI, CPRA, DCIA, GDPR, HIPAA, PDP, SOX y más) que requieren que las empresas tomen medidas preventivas para proteger los datos confidenciales de las personas.

Igualmente importante es que, incluso después de que los datos se anonimicen, se pueden seguir utilizando para fines de análisis, información comercial, toma de decisiones e investigación, sin revelar nunca la información personal de nadie.

1.8.3 Tipos de anonimización de datos

Existen 6 tipos básicos de anonimización de datos, entre ellos:

1. **Enmascaramiento de datos:** El software de enmascaramiento de datos reemplaza datos confidenciales, como números de tarjetas de crédito, números de licencia de conducir y números de la Seguridad Social, con caracteres, dígitos o símbolos sin sentido, o datos enmascarados aparentemente realistas, pero ficticios. El enmascaramiento de datos de prueba los hace disponibles para fines de desarrollo o prueba, sin comprometer la privacidad de la información original. El enmascaramiento de datos se puede aplicar a un campo específico o a conjuntos de datos completos, utilizando una variedad de técnicas como la sustitución de caracteres, la mezcla de datos y el truncamiento. Los datos se pueden enmascarar a pedido o según un cronograma. El conjunto de enmascaramiento de datos incluye la tokenización de datos, que sustituye irreversiblemente los datos personales con marcadores de posición aleatorios, y la generación de datos sintéticos, cuando la cantidad de datos de producción es insuficiente.
2. **Seudonimización:** la seudonimización anonimiza los datos reemplazando cualquier información de identificación con un identificador seudónimo o pseudónimo. La información personal

que se reemplaza comúnmente incluye nombres, direcciones y números de la Seguridad Social. Los datos seudonimizados reducen el riesgo de exposición o uso indebido de información personal identificable, al mismo tiempo que permiten que el conjunto de datos se use con fines legítimos. En la ecuación de seudonimización vs. anonimización, la primera es reversible (a diferencia de las soluciones de tokenización de datos) y, a menudo, se usa en combinación con otras tecnologías que mejoran la privacidad, como el enmascaramiento de datos vs. el cifrado.

3. **Agregación de datos:** la agregación de datos, que combina datos recopilados de muchas fuentes diferentes en una sola vista, se usa para obtener información para una mejor toma de decisiones o análisis de tendencias y patrones. Los datos se pueden agregar en diferentes niveles de granularidad, desde simples resúmenes hasta cálculos complejos, y se puede hacer en datos categóricos, datos numéricos y datos de texto. Los datos agregados se pueden presentar en diversas formas y se pueden utilizar para diversos fines, como análisis, informes y visualización. También se puede realizar con datos que se han seudonimizado o enmascarado para proteger aún más la privacidad individual.

4. **Generación aleatoria de datos:** La generación aleatoria de datos, que mezcla aleatoriamente los datos para ocultar información confidencial, se puede aplicar a un conjunto de datos completo o a campos o columnas específicos de una base de datos. La generación aleatoria de datos, que suele utilizarse junto con herramientas de enmascaramiento de datos o tokenización de datos, es ideal para ensayos clínicos, ya que garantiza que los sujetos no solo se elijan al azar, sino que también se asignen aleatoriamente a diferentes grupos de tratamiento. Al combinar diferentes tipos de anonimización de datos, se reduce el sesgo y se aumenta la validez de los resultados.

5. **Generalización de datos:** La generalización de datos, que reemplaza valores de datos específicos por valores más generalizados, se utiliza para ocultar información de identificación personal (PII), como direcciones o edades, a terceros no autorizados. Sustituye categorías, rangos o áreas geográficas por valores específicos. Por ejemplo, una dirección específica, como 1705 Fifth Avenue, se puede generalizar al centro, al centro de la ciudad o a la zona alta de la ciudad. De manera similar, la edad de 55 años se puede generalizar a un grupo de edad llamado de 50 a 60 años, o adultos de mediana edad.

6. **Intercambio de datos:** El intercambio de datos reemplaza los valores de datos reales por otros ficticios, pero similares. Por ejemplo, un nombre real, como Don Johnson, se puede intercambiar por uno ficticio, como Robbie Simons. O una dirección real, como 186 South Street, se puede intercambiar por una ficticia, como 15 Parkside Lane. El intercambio de datos es similar al generador de datos aleatorios, pero en lugar de mezclar los datos, reemplaza los valores originales por otros nuevos y ficticios.

1.8.4 Técnicas de anonimización de datos

Existen 5 técnicas clave de anonimización de datos, entre ellas:

1. **Anonimato K:** El anonimato K garantiza que la información de ninguna persona pueda distinguirse de al menos "K-1" otras personas en el mismo conjunto de datos. En otras palabras, para cualquier registro dado, hay al menos K otros registros en el conjunto de datos con valores idénticos para todos los atributos de identificación. Por ejemplo, si un conjunto de datos contiene información personal como nombres, direcciones y números de seguro social, y K se establece en 3, entonces la información de ninguna persona puede distinguirse de al

menos otras 2 en el conjunto de datos. Esto significa que los piratas informáticos no podrán identificar a una persona específica dentro del conjunto de datos simplemente mirando los valores de los atributos de identificación, porque hay al menos otras 2 personas en el conjunto de datos con exactamente los mismos valores. El anonimato K nunca puede garantizar una protección de la privacidad del 100%, porque a medida que aumenta el valor de K, el riesgo de reidentificación disminuye, pero nunca se elimina por completo. Además, esta técnica de anonimización de datos no tiene en cuenta ningún factor externo a la hora de identificar a alguien, por lo que incluso cuando un conjunto de datos es K-anónimo, puede combinarse con otras fuentes de datos para volver a identificar a una persona específica.

2. **L Diversity**, que garantiza que la información de ninguna persona pueda distinguirse de al menos L otras personas del conjunto de datos en función de un atributo sensible, es una extensión de K Anonymity. Pero mientras que K Anonymity garantiza que la información de ninguna persona pueda distinguirse de al menos K-1 otras personas del conjunto de datos, L Diversity protege los atributos sensibles, así como los generales. Por ejemplo, si un conjunto de datos contiene atributos sensibles como una condición médica o medicamentos recetados, debe haber al menos L personas en ese conjunto de datos para cualquier valor específico del atributo sensible, a fin de no identificar a una persona específica. Al igual que K Anonymity, L Diversity no garantiza una protección total de la privacidad, por las mismas razones citadas en la sección anterior. Y la diversidad L es más difícil de implementar que el anonimato K, porque no solo tiene que identificar y proteger atributos sensibles, sino que solo puede funcionar cuando al menos L valores distintos para cada uno de esos atributos están presentes en el conjunto de datos.

3. **T Closeness:** T Closeness contribuye a la eficacia de la combinación de anonimato K / diversidad L al asegurar que la distribución de los atributos sensibles en el conjunto de datos coincida con la de la población objetivo, lo más fielmente posible. Por ejemplo, si un conjunto de datos determinado contiene no solo información personal identificable, sino también atributos sensibles como el ingreso, T Closeness garantiza que la distribución del ingreso en el conjunto de datos sea muy cercana a la de la población objetivo. De esa manera, el valor del ingreso no revela ninguna información sobre una persona en particular. Al igual que el anonimato K y la diversidad L, T Closeness no puede garantizar una protección completa de la privacidad, por las mismas razones citadas anteriormente. Y la cercanía T es incluso más difícil de implementar que el anonimato K o la diversidad L, porque no solo tiene que identificar y proteger atributos sensibles, sino que solo puede ser efectiva cuando la distribución de los atributos sensibles en el conjunto de datos es similar a la de la población.

4. **Privacidad diferencial:** La privacidad diferencial, que añade ruido aleatorio a los datos para que no sean identificables, es un marco matemático utilizado en el análisis, la elaboración de informes y la visualización de datos que busca equilibrar el riesgo de privacidad de un conjunto de datos determinado frente a su utilidad. Utiliza varias técnicas de aleatorización, como la perturbación y el muestreo. Un parámetro de nivel de protección de la privacidad, conocido como épsilon ( ), controla la cantidad de ruido añadido a los datos. Cuanto menor sea el valor de épsilon, mayor será el nivel de ruido necesario. La privacidad diferencial puede hacer que los datos sean menos precisos, por lo que es importante encontrar el equilibrio adecuado entre la protección de la privacidad y la utilidad. Y como siempre hay una pequeña probabilidad de reidentificación (controlada por el parámetro de privacidad), no puede garantizar una protección completa.

5. **Respuesta aleatoria:** La respuesta aleatoria es una técnica de encuesta que funciona al

decidir aleatoriamente si una pregunta se responde con sinceridad o si se da una respuesta predeterminada de Sí o No. Permite a las personas responder con sinceridad a preguntas delicadas, sin revelar sus respuestas reales. Esto se logra introduciendo un nivel de aleatoriedad en el proceso de encuesta, con el fin de evitar que los administradores de la encuesta conozcan la respuesta verdadera. En una encuesta sobre el consumo de drogas, por ejemplo, una de las preguntas podría ser "¿Alguna vez ha consumido drogas ilegales?". Esta técnica asigna aleatoriamente a cada encuestado la opción de responder honestamente o dar una respuesta predeterminada de "Sí" con una cierta probabilidad (digamos 0,5). La técnica de respuesta aleatoria se puede combinar con otros métodos de encuesta, como encuestas anónimas y encuestas autoadministradas, para proteger aún más la privacidad de los encuestados. Como concepto probabilístico, la respuesta aleatoria no puede brindar una protección integral de la privacidad, porque la reidentificación es posible, aunque sea de forma remota.

1.8.5 Anonimización de datos:

A continuación, se incluye un resumen de las ventajas y desventajas de la anonimización de datos:

Pros	Contras
Hace que la identificación de una persona en un conjunto de datos sea imposible o muy improbable	Puede reducir la utilidad de los datos al modificar o eliminar elementos PII importantes
Permite compartir datos con fines legítimos, como análisis e investigación	Puede permitir la reidentificación, si un atacante puede hacer referencias cruzadas de datos adicionales
Permite un cumplimiento más rápido y sencillo de las leyes de privacidad de datos	Puede requerir experiencia y herramientas especializadas, lo que aumenta la complejidad y el costo
Impide que los atacantes obtengan acceso a información confidencial	Puede no proporcionar protección total de la privacidad de los datos (si la reidentificación tiene éxito)
Minimiza el riesgo de errores, como la vinculación incorrecta de datos	Puede no funcionar con datos muy confidenciales o que tienen propiedades únicas
Reduce los costos, con la reutilización de datos sin consentimiento y sin necesidad de almacenamiento seguro	Puede consumir mucho tiempo y recursos y no es muy escalable

1.9 1.1.2 Tipos de métodos de recopilación de datos

Ahora, exploremos los diferentes métodos de recopilación de datos con mayor detalle, incluidos ejemplos de cuándo y cómo usarlos de manera efectiva:

1.9.1 Encuestas y cuestionarios

Las encuestas y los cuestionarios son herramientas versátiles para recopilar datos de una gran cantidad de encuestados. Se utilizan comúnmente para:

- **Comentarios de clientes:** Recopilar opiniones y comentarios sobre productos, servicios y satisfacción general.

- **Investigación de mercado:** Evaluar las preferencias del mercado, identificar tendencias y evaluar el comportamiento del consumidor.
- **Encuestas a empleados:** Medir el compromiso de los empleados, la satisfacción laboral y los comentarios sobre las condiciones del lugar de trabajo.

Ejemplo: si tiene un negocio de comercio electrónico y desea comprender las preferencias de los clientes, puede crear una encuesta en línea para preguntarles sobre sus categorías de productos favoritas, métodos de pago preferidos y frecuencia de compra.

1.9.2 Entrevistas

Las entrevistas implican conversaciones individuales o grupales con los participantes para recopilar información detallada. Son particularmente útiles para:

- **Investigación cualitativa:** Explorar temas complejos, motivaciones y experiencias personales.
- **Análisis en profundidad:** Obtener una comprensión profunda de problemas o situaciones específicas.
- **Opiniones de expertos:** Entrevistar a expertos de la industria o líderes de opinión para recopilar información valiosa.

Ejemplo: si eres un proveedor de atención médica que busca mejorar las experiencias de los pacientes, realizar entrevistas con los pacientes puede ayudarte a descubrir puntos críticos específicos y sugerencias para mejorar.

1.9.3 Observaciones o estudios de casos

Las observaciones implican observar y registrar comportamientos o eventos en su contexto natural. Este método es ideal para:

- **Estudios de comportamiento** Analizar cómo las personas interactúan con productos o entornos.
- **Investigación de campo:** Recopilar datos en entornos del mundo real, como tiendas minoristas, espacios públicos o aulas.
- **Investigación etnográfica:** Sumergirte en una cultura o comunidad específica para comprender sus prácticas y costumbres.

Ejemplo: si administras una tienda minorista, observar el flujo de tráfico de clientes y los comportamientos de compra puede ayudar a optimizar el diseño de la tienda y la ubicación de los productos.

1.9.4 Herramientas de recopilación de datos en línea

La mayoría de los métodos de recopilación de datos enumerados anteriormente también se pueden utilizar para recopilar datos del mundo en línea, especialmente si busca recopilar datos cualitativos de los usuarios de Internet. Otras herramientas en línea incluyen:

- **Sistemas de gestión de la información:** Aunque generalmente están diseñados para administrar su base de datos, estos sistemas de gestión también pueden ayudarlo a recopilar datos, especialmente datos internos generados por su organización.
- **Software de recopilación de datos:** Existe otro software de recopilación de datos que le facilita la recopilación de datos de Internet y de los usuarios de Internet. Por ejemplo, Google

Forms le permite crear formularios. Esto se puede utilizar para crear formularios de solicitud de empleo, lo que permite recopilar fácilmente los datos de los solicitantes.

1.9.5 Análisis de documentos

El análisis de documentos implica revisar y extraer información de documentos escritos o digitales. Es valioso para:

- **Investigación histórica:** Estudiar registros históricos, manuscritos y archivos.
- **Análisis de contenido:** Analizar contenido textual o visual de sitios web, informes o publicaciones.
- **Legal y cumplimiento:** Revisión de contratos, políticas y documentos legales para fines de cumplimiento.

Ejemplo: si eres un comercializador de contenido, puedes analizar las publicaciones de blogs de la competencia para identificar temas y palabras clave comunes que se usan en tu industria.

1.9.6 Web Scraping

El web scraping es el proceso automatizado de extracción de datos de sitios web. Es adecuado para:

- **Análisis de la competencia:** Recopilación de datos sobre precios de productos de la competencia, descripciones y opiniones de clientes.
- **Investigación de mercado:** Recopilación de datos sobre listados de productos, reseñas y tendencias de sitios web de comercio electrónico.
- **Seguimiento de noticias y redes sociales:** Seguimiento de artículos de noticias, publicaciones en redes sociales y comentarios relacionados con su marca o industria.

El web scraping es un proceso en el que se utilizan bots automatizados para rastrear Internet y extraer datos. Los bots recopilan información descomponiendo primero el sitio de destino en su forma más básica, texto HTML, y luego escanean para recopilar datos de acuerdo con algunos parámetros preestablecidos. Después de eso, los datos recopilados se entregan en formato CSV o Excel, por lo que son legibles para quien quiera usarlos. Los web scrapers se encuentran entre los métodos más eficientes que puede emplear.

Ejemplo: si trabaja en el sector de viajes, el web scraping puede ayudarle a recopilar datos de precios de vuelos y alojamiento de varios sitios web de reserva de viajes para mantenerse competitivo.

1.9.7 Monitoreo de redes sociales

El monitoreo de redes sociales implica el seguimiento y análisis de conversaciones y actividades en plataformas de redes sociales. Es valioso para:

- **Gestión de reputación de marca:** Monitoreo de menciones y sentimientos de marca para abordar las inquietudes de los clientes o capitalizar los comentarios positivos.
- **Análisis de la competencia:** Controlar las estrategias de redes sociales de la competencia y la interacción con los clientes.
- **Identificación de tendencias:** Identificar tendencias emergentes y contenido viral dentro de su industria.

Ejemplo: si tiene un restaurante, el monitoreo de redes sociales puede ayudarlo a realizar un

seguimiento de las reseñas, comentarios y hashtags de los clientes relacionados con su establecimiento, lo que le permite responder rápidamente a los comentarios y tendencias de los clientes.

Al comprender los matices y las aplicaciones de estos métodos de recopilación de datos, puede elegir el enfoque más adecuado para recopilar información valiosa para sus objetivos específicos. Recuerde que una estrategia de recopilación de datos bien pensada es la piedra angular de la toma de decisiones informada y el éxito comercial.

Una vez que tenga en sus manos la herramienta adecuada para la recopilación de datos en línea, comenzará a ver lo que se ha perdido y lo indispensables que son estas herramientas. Ofrecen numerosos beneficios a empresas de todo tipo. Veamos algunos de ellos.

- **Análisis de rendimiento** La recopilación de datos a través de una herramienta fiable resuelve muchos problemas. Después de recopilar datos fiables, mejora el rendimiento de su negocio. Esto se debe a que la mayoría de las acciones de su negocio se toman en función de los datos que ha recopilado. La calidad y la precisión de los datos utilizados influyen en gran medida en el resultado general de su negocio.
- **Entender a los clientes** Sin datos, es apropiado decir que su negocio está funcionando a ciegas. Esto se debe a que no tendrá ninguna información sobre su mercado y sus clientes. Ninguna empresa quiere operar así. La recopilación de datos resulta útil, especialmente ahora que Internet es una mina de oro de información. Las redes sociales ofrecen muchas oportunidades en términos de recopilación de datos. Con la herramienta adecuada para recopilar suficientes datos, puede comprender a sus clientes y al mercado, lo que afecta positivamente al crecimiento de su negocio.
- **Encontrar soluciones** a los problemas comerciales ¿Una campaña salió mal? ¿O un contenido publicado no está obteniendo suficiente tracción? Puedes implementar herramientas de scraping para ayudarte a analizar tus canales de marketing y el tipo de recepción que está recibiendo tu contenido. En este punto, los problemas se pueden analizar y solucionar posteriormente.

1.10 ¿Cómo diseñar sus instrumentos de recolección de datos?

Ahora que ha definido sus preguntas de investigación, identificado las fuentes de datos, establecido objetivos claros y elegido los métodos de recolección de datos adecuados, es hora de diseñar los instrumentos que utilizará para recolectar datos de manera eficaz.

1.10.1 Diseñe preguntas de encuesta eficaces

Diseñar preguntas de encuesta es un paso crucial para recopilar datos precisos y significativos. A continuación, se incluyen algunas consideraciones clave:

- **Claridad:** Asegúrese de que sus preguntas sean claras y concisas. Evite la jerga o el lenguaje ambiguo que pueda confundir a los encuestados.
- **Relevancia:** Formule preguntas que se relacionen directamente con sus objetivos de investigación. Evite preguntas innecesarias o irrelevantes que puedan generar fatiga en la encuesta.
- **Evite las preguntas capciosas:** Formule preguntas que no guíen a los encuestados hacia una respuesta en particular. Mantenga la neutralidad para obtener respuestas imparciales.
- **Opciones de respuesta:** Proporcione opciones de respuesta adecuadas, incluidas opciones múltiples, escalas de Likert o formatos abiertos, según el tipo de datos que necesite.

- **Prueba piloto:** Antes de implementar su encuesta, realice pruebas piloto con un grupo pequeño para identificar cualquier problema con la redacción de las preguntas o las opciones de respuesta.

1.10.2 Cree preguntas de entrevista para conversaciones reveladoras

El desarrollo de preguntas de entrevista requiere una consideración cuidadosa para obtener información valiosa de los participantes:

- **Preguntas abiertas:** Utilice preguntas abiertas para alentar a los participantes a compartir sus pensamientos, experiencias y perspectivas sin verse limitados por respuestas predefinidas.
- **Preguntas de sondeo:** Prepare preguntas de seguimiento para profundizar en temas específicos o aclarar respuestas.
- **Entrevistas estructuradas vs. semiestructuradas:** Decida si sus entrevistas seguirán un formato estructurado con preguntas predefinidas o un enfoque semiestructurado que permita flexibilidad.
- **Evite las preguntas sesgadas:** Asegúrese de que sus preguntas no dirijan a los participantes hacia las respuestas deseadas. Mantenga la objetividad durante toda la entrevista.

1.10.3 Cree una lista de verificación de observación para la recopilación de datos

Al realizar observaciones, es fundamental contar con una lista de verificación bien estructurada:

- **Variables claramente definidas:** Identifique las variables o comportamientos específicos que está observando y asegúrese de que estén bien definidos.
- **Formato de lista de verificación:** Cree un formato de lista de verificación que sea fácil de usar y seguir durante las observaciones. Esto puede incluir casillas de verificación, escalas o espacio para notas.
- **Capacitación de observadores:** Si tiene un equipo de observadores, proporcióneles una capacitación exhaustiva para garantizar la coherencia y precisión en la recopilación de datos.
- **Observaciones piloto:** Antes de comenzar la recopilación formal de datos, realice observaciones piloto para refinar su lista de verificación y asegurarse de que capture la información necesaria.

1.10.4 Agilice la recopilación de datos con formularios y plantillas

La creación de formularios y plantillas de recopilación de datos fáciles de usar ayuda a agilizar el proceso:

- **Coherencia:** Asegúrese de que todos los formularios de recopilación de datos sigan un formato y una estructura coherentes, lo que facilitará la comparación y el análisis de los datos.
- **Validación de datos:** Incorpore comprobaciones de validación de datos para reducir los errores durante la entrada de datos. Esto puede incluir menús desplegables, selectores de fechas o campos obligatorios.
- **Formularios digitales o en papel:** Decida si los formularios digitales o los formularios en papel tradicionales son más adecuados para sus necesidades de recopilación de datos. Los formularios digitales suelen ofrecer validación de datos en tiempo real y acceso remoto.
- **Accesibilidad:** Asegúrese de que sus formularios y plantillas sean accesibles para todos los miembros del equipo que participan en la recopilación de datos. Brinde capacitación si es necesario.

1.10.5 Grupos de discusión

Los grupos de discusión son debates guiados entre personas seleccionadas para obtener información sobre sus puntos de vista y experiencias.

- **Ventajas:** Los grupos de discusión permiten la interacción entre los participantes, lo que puede generar una amplia gama de opiniones e ideas. Son buenos para explorar nuevos temas sobre los que hay poco conocimiento previo.
- **Desventajas:** Las voces dominantes en el grupo pueden influir en la discusión, lo que podría silenciar a los participantes menos asertivos. También requieren facilitadores capacitados para moderar la discusión de manera efectiva.

1.11 1.1.3 Agregar datos de múltiples fuentes e integrarlos en conjuntos de datos.

1.11.1 Definición y concepto de integración de datos

Integración de datos se refiere al proceso de combinar datos de múltiples fuentes en una vista unificada y coherente. Implica reunir datos que residen en diferentes sistemas, formatos o bases de datos y transformarlos en un formato coherente y significativo. El objetivo es crear una vista unificada y completa de los datos, lo que permite a las organizaciones obtener información valiosa y tomar decisiones informadas.

1.11.2 Enfoques y técnicas comunes de integración de datos

Procesos ETL (Extracción, Transformación, Carga) ETL es un enfoque tradicional de la integración de datos que implica extraer datos de los sistemas de origen, transformarlos para cumplir con los requisitos del sistema de destino y cargarlos en un sistema de destino. Este proceso normalmente implica la limpieza, agregación, filtrado y formateo de datos.

La virtualización de datos permite a las organizaciones acceder e integrar datos de múltiples fuentes sin moverlos ni replicarlos físicamente. Proporciona una capa virtual que abstrae las fuentes de datos subyacentes, lo que permite a los usuarios consultar y recuperar datos como si estuvieran almacenados en una única ubicación.

Consideraciones para seleccionar una estrategia de integración de datos adecuada

Al elegir una estrategia de integración de datos, se deben considerar varios factores

Considere el tamaño y la velocidad a la que se generan los datos y se deben integrar. La integración en tiempo real puede ser necesaria para flujos de datos de alta velocidad, mientras que el procesamiento por lotes puede ser suficiente para datos menos sensibles al tiempo.

Evalúe la complejidad de las fuentes de datos, incluidas las variaciones en los formatos, las estructuras y la semántica de los datos. Algunos enfoques de integración pueden ser más adecuados que otros para gestionar fuentes de datos complejas.

1.11.3 Preparación de datos para la integración

Evaluación de la calidad de los datos y elaboración de perfiles de datos

La elaboración de perfiles de datos implica examinar la estructura, el contenido y la calidad de los datos antes de integrarlos. Ayuda a identificar los tipos de datos, los formatos y los problemas potenciales.

La evaluación de la calidad de los datos implica evaluar la precisión, la integridad, la coherencia y la validez de los datos. Este paso ayuda a identificar los problemas de calidad de los datos que deben abordarse.

Limpieza y estandarización de datos

La limpieza de datos implica eliminar o corregir errores, inconsistencias y duplicados en los datos. Garantiza la integridad de los datos y mejora la calidad general de los datos integrados.

La estandarización de datos implica transformar los datos a un formato o estructura común, haciéndolos consistentes en diferentes fuentes. Este paso ayuda a facilitar la integración de datos al resolver las disparidades de formato y estructura.

Mapeo de datos y alineación de esquemas

El mapeo de datos implica hacer coincidir y vincular elementos de datos de diferentes fuentes en función de su semántica o relaciones. Define cómo se integrarán y alinearán los datos de varias fuentes.

La alineación de esquemas garantiza que los esquemas o estructuras de datos de distintas fuentes sean compatibles y se puedan integrar sin problemas. Implica la asignación y conciliación de diferencias en los nombres de atributos, los tipos de datos y las relaciones.

1.11.4 Conexión y recopilación de datos de distintas fuentes

Identificación de fuentes de datos relevantes

Comience por identificar las distintas fuentes de datos que son relevantes para su organización o proyecto. Esto podría incluir bases de datos, aplicaciones, servicios en la nube, plataformas de redes sociales, dispositivos IoT o proveedores de datos externos. Determine los tipos de datos que proporciona cada fuente y el valor potencial que pueden ofrecer en términos de información o toma de decisiones.

Métodos de extracción de datos:

Conexiones directas a bases de datos:

Establezca conexiones a las bases de datos directamente mediante protocolos adecuados como ODBC (Open Database Connectivity) o JDBC (Java Database Connectivity). Extraiga datos mediante consultas SQL u otros métodos de extracción específicos de la base de datos.

API (Application Programming Interfaces):

Muchas aplicaciones y servicios proporcionan API que permiten el acceso programático a sus datos. Identifique las API que están disponibles para las fuentes de datos deseadas y aprenda a autenticar, solicitar datos y manejar respuestas.

Transmisión de datos e integración de datos en tiempo real:

En determinados escenarios, se requiere la integración de datos en tiempo real o casi en tiempo real. Implemente tecnologías de transmisión de datos como Apache Kafka, AWS Kinesis o Azure

Event Hubs para ingerir y procesar datos a medida que se generan. Establezca flujos de trabajo y canalizaciones de datos para recopilar e integrar continuamente datos de transmisión. La recopilación de datos de diferentes fuentes requiere una planificación cuidadosa y la consideración de los siguientes factores:

Compatibilidad de fuentes de datos

Asegúrese de que los métodos de extracción de datos sean compatibles con las fuentes de datos de las que desea recopilarlos. Considere la disponibilidad y accesibilidad de las API, los conectores o los formatos de archivo para cada fuente de datos.

Autenticación y autorización

Comprenda los mecanismos de autenticación necesarios para acceder y recopilar datos de cada fuente. Implemente los protocolos de autenticación necesarios, como claves API, OAuth o autenticación basada en tokens.

Volumen de datos y escalabilidad

Considere el volumen de datos que necesita recopilar y asegúrese de que sus procesos de recopilación de datos puedan manejar grandes cantidades de datos de manera eficiente. Implementar soluciones escalables para adaptarse a volúmenes de datos crecientes o picos repentinos de datos.

1.11.5 Transformación y armonización de datos

Comprensión de los formatos y estructuras de los datos

Antes de integrar los datos, es importante comprender claramente los distintos formatos y estructuras de datos presentes en las distintas fuentes. Esto incluye reconocer las diferencias en los tipos de datos, las longitudes de los campos, los esquemas de codificación y otras variaciones estructurales. Comprender estos matices ayuda a desarrollar estrategias de transformación adecuadas.

Técnicas de transformación de datos

La transformación de datos implica manipular los datos para alinearlos con un formato o estructura común. Se pueden emplear varias técnicas para este propósito:

Agregación y resumen de datos Esto implica consolidar datos de múltiples fuentes en un formato unificado agregando puntos de datos similares o resumiéndolos según criterios específicos.

Garantizar la coherencia e integridad de los datos durante la transformación

Durante el proceso de transformación, mantener la coherencia e integridad de los datos es primordial. Implica realizar la validación de datos, lo que garantiza que los datos transformados cumplan con las reglas y restricciones predefinidas. Al aplicar reglas de negocios y técnicas de validación de datos, como referencias cruzadas y conciliación, se pueden identificar y resolver inconsistencias y errores.

Validación de datos y control de calidad

La validación de datos y el control de calidad son pasos cruciales en el proceso de integración y aceptación de datos de diferentes fuentes. Esta etapa garantiza que los datos integrados sean precisos, completos, consistentes y confiables para su posterior análisis y toma de decisiones. A continuación, se ofrecen algunas explicaciones de los aspectos clave involucrados en la validación de datos y el control de calidad:

Técnicas de validación de datos

La validación de datos implica la verificación de la integridad de los datos y el cumplimiento de reglas y restricciones predefinidas. Se pueden utilizar varias técnicas para validar los datos, entre ellas:

- **Perfiles de datos y análisis estadístico**

  Los perfiles de datos examinan las características y propiedades de los datos para identificar anomalías, como valores faltantes, valores atípicos o formatos inconsistentes. El análisis estadístico ayuda a identificar patrones y tendencias, lo que garantiza que los datos se alineen con las distribuciones y relaciones esperadas.

Estrategias de mejora y evaluación de la calidad de los datos

La evaluación de la calidad de los datos evalúa la calidad general de los datos integrados. Esta evaluación implica examinar varias dimensiones de la calidad de los datos, como la precisión, la integridad, la coherencia, la puntualidad y la relevancia. Las estrategias para mejorar la calidad de los datos pueden incluir:

- **Limpieza de datos**

  La limpieza de datos implica identificar y corregir o eliminar errores, inconsistencias y duplicaciones dentro del conjunto de datos integrado. Puede incluir técnicas como estandarización de datos, enriquecimiento de datos y eliminación de datos redundantes o irrelevantes.

Abordar inconsistencias y errores de datos

La integración de datos a menudo implica combinar datos de fuentes dispares, lo que puede generar inconsistencias y errores. Abordar estos problemas requiere un análisis y una resolución cuidadosos:

- **Conciliación de datos**

  Las inconsistencias identificadas durante el proceso de referencia cruzada y conciliación deben resolverse alineando y armonizando los datos de diferentes fuentes. Esto puede implicar transformaciones de datos, estandarización o mapeo de valores de datos.

  La integración y aceptación de datos de diferentes fuentes es un proceso crítico en el mundo actual impulsado por los datos. Esta guía completa ha cubierto varios aspectos de la integración de datos, destacando su importancia, desafíos y mejores prácticas. Al integrar con éxito datos de diversas fuentes, las organizaciones pueden descubrir información valiosa, tomar decisiones informadas y obtener una ventaja competitiva.

1.12 1.1.4 Explicar las distintas soluciones de almacenamiento de datos

1.12.1 ¿Qué es el almacenamiento de datos?

El almacenamiento de datos es la retención de información mediante tecnología diseñada explícitamente para almacenar esos datos y hacerlos lo más accesibles posible. Las formas más frecuentes de almacenamiento de datos son el almacenamiento en archivos, en bloques y en objetos, cada uno ideal para diferentes propósitos.

Los usuarios proporcionan datos de entrada, pero las computadoras necesitan una forma de almacenarlos más allá de la memoria de acceso aleatorio (RAM) temporal. Si bien la memoria de solo lectura (ROM) ofrece almacenamiento permanente, no se puede editar.

El desafío era encontrar una solución rentable y espaciosa que retuviera los datos incluso después de los apagados y permitiera realizar cambios. La solución: el almacenamiento de datos.

Las empresas impulsadas por los datos, en particular, dependen en gran medida de encontrar soluciones confiables y efectivas para almacenar sus grandes cantidades de datos.

Esto incluye el almacenamiento tradicional de archivos y bloques, pero también las soluciones de almacenamiento de objetos, cada vez más importantes, para administrar cantidades masivas de datos no estructurados, como imágenes, videos y archivos.

1.12.2 ¿Por qué es importante el almacenamiento de datos?

Para abordar las demandas informáticas de alto nivel actuales, como los proyectos de big data, la inteligencia artificial (IA), el aprendizaje automático (ML) y el Internet de las cosas (IoT), las empresas y las personas necesitan almacenamiento de datos. La otra cara de la necesidad de un almacenamiento masivo de datos es la protección contra la pérdida de datos debido a desastres, fallas o fraudes. Las organizaciones también pueden utilizar el almacenamiento de datos como una opción de respaldo para evitar la pérdida de datos.

1.12.3 Formas de almacenamiento de datos

Los datos se pueden recopilar y almacenar de tres formas: archivos, bloques u objetos.

1. **Almacenamiento de archivos**, también conocido como almacenamiento a nivel de archivo o almacenamiento basado en archivos, es un sistema de almacenamiento jerárquico para organizar y almacenar datos. Los datos se guardan en archivos, luego se organizan en carpetas y se estructuran en una jerarquía de directorios y subdirectorios.
2. **Almacenamiento en bloques** es una tecnología que se utiliza para almacenar datos en bloques. Luego, los bloques se guardan por separado, cada uno con su propia identidad única. Los desarrolladores utilizan el almacenamiento en bloques para configuraciones de computadora que requieren un transporte de datos rápido, eficiente y confiable.
3. **Almacenamiento de objetos** es una arquitectura diseñada para manejar volúmenes masivos de datos no estructurados. Estos datos no caben, o no se pueden estructurar en, una base de datos relacional estándar con filas y columnas. Algunos ejemplos incluyen correo electrónico, películas, imágenes, páginas web, archivos de audio, datos de sensores y contenido multimedia y en línea (textual o no textual).

1.12.4 Tipos de almacenamiento de datos

Los usuarios necesitan dispositivos de almacenamiento para almacenar datos en cualquier formato. Los dispositivos de almacenamiento de datos se clasifican en almacenamiento de área directa y almacenamiento basado en red.

Almacenamiento de área directa Como indica el nombre, el almacenamiento de área directa o de conexión directa (DAS) suele estar cerca y conectado directamente al equipo informático utilizado. A menudo, es la única máquina conectada a él. El DAS también puede proporcionar servicios de copia de seguridad locales adecuados, aunque el uso compartido está restringido.

Los disquetes, los discos ópticos o los discos compactos (CD), los discos de vídeo digital (DVD), las unidades de disco duro (HDD), las unidades flash y las unidades de estado sólido (SSD) son

ejemplos de dispositivos DAS.

Almacenamiento basado en red El almacenamiento basado en red permite que varias computadoras accedan a él a través de una red, lo que lo hace ideal para compartir datos y colaborar. Su capacidad para almacenar datos fuera del sitio lo hace ideal para copias de seguridad de bases de datos y seguridad de datos.

Network-attached storage (NAS) y storage area network (SAN) son dos configuraciones típicas de almacenamiento basado en red.

NAS suele ser un dispositivo único con una matriz redundante de unidades independientes (RAID). El almacenamiento SAN puede definirse como una red de muchos dispositivos, como SSD y almacenamiento flash, almacenamiento híbrido, almacenamiento híbrido en la nube, software y dispositivos de respaldo y almacenamiento de datos en la nube.

Dispositivos de almacenamiento de datos Numerosos sistemas de almacenamiento de datos ofrecen seguridad de información confiable. La memoria de la computadora y el almacenamiento local pueden no ser suficientes para preservar los datos privados. El almacenamiento de datos no volátil es la mejor opción de seguridad, que no requiere energía constante para almacenar y retener datos.

Almacenamiento en SSD y memoria flash El almacenamiento flash es una tecnología de estado sólido que escribe y almacena datos mediante chips de memoria flash. Una unidad flash de disco de estado sólido (SSD) utiliza memoria flash para almacenar datos.

En comparación con las unidades de disco duro (HDD), un sistema de estado sólido no tiene componentes móviles y, por lo tanto, reduce la latencia, lo que requiere menos SSD. Debido a que la mayoría de los SSD actuales están basados en flash, el almacenamiento flash es sinónimo de almacenamiento de estado sólido.

Almacenamiento híbrido Los SSD y la memoria flash tienen un rendimiento más rápido que los HDD, aunque las matrices completamente flash pueden ser costosas. Muchas empresas utilizan un método híbrido que combina la velocidad flash con la capacidad de almacenamiento de los discos duros.

Una infraestructura de almacenamiento bien equilibrada permite a las empresas elegir la tecnología adecuada para diversos requisitos de almacenamiento. Proporciona una alternativa rentable a la migración de los discos duros tradicionales a las unidades flash.

Almacenamiento en la nube El almacenamiento en la nube es más rentable y escalable que mantener el contenido en las instalaciones en discos duros o redes de almacenamiento. Los proveedores de servicios en la nube ayudan a almacenar datos y archivos en un lugar remoto accesible a través de Internet público o una conexión de red privada dedicada.

El proveedor aloja, protege, administra y mantiene los servidores y la infraestructura relacionada, lo que garantiza la accesibilidad siempre que sea necesario. Por eso, las empresas que buscan mejorar las capacidades organizativas, operativas y técnicas están migrando las cargas de trabajo y los centros de datos locales a la nube.

Consejo: el software de migración a la nube reemplaza el hardware obsoleto, elimina las actualizaciones costosas y pone fin a los alquileres de centros de datos de alto precio.

Almacenamiento en la nube híbrido El almacenamiento en la nube híbrido incorpora componentes de las nubes públicas y privadas. Las organizaciones pueden elegir en qué nube almacenar los

datos utilizando el almacenamiento en la nube híbrido. Por ejemplo, los datos altamente regulados que requieren un archivado y una replicación estrictos generalmente se adaptan mejor a un entorno de nube privada.

Por otro lado, los datos menos confidenciales se pueden guardar en la nube pública. Algunas empresas utilizan nubes híbridas para complementar sus redes de almacenamiento internas con almacenamiento en la nube pública.

1.12.5 Software y aplicaciones de respaldo

El almacenamiento y los dispositivos de respaldo protegen contra la pérdida de datos debido a catástrofes, fallas o fraudes. Crean copias de seguridad periódicas de los datos y las aplicaciones en un dispositivo secundario diferente, que luego utilizan para la recuperación ante desastres.

Los dispositivos de respaldo van desde unidades de disco duro y unidades de estado sólido hasta unidades de cinta y servidores, pero el almacenamiento de respaldo también se puede proporcionar como un servicio, a menudo conocido como respaldo como servicio (BaaS). BaaS, al igual que otras soluciones como servicio, ofrece una alternativa de bajo costo para la protección de datos al almacenarlos en un lugar distante con escalabilidad.

1.12.6 Beneficios de un almacenamiento de datos eficiente

Además de ser más rápido y más confiable que las soluciones de almacenamiento en papel, el almacenamiento de datos digitales ofrece una serie de otras ventajas.

- Conservación de datos a largo plazo. El almacenamiento de datos digitales facilita la recopilación de enormes cantidades de información durante períodos prolongados.
- Acceso más fácil. En lugar de ir físicamente a una habitación llena de archivadores, todos pueden recuperar de inmediato la información que necesitan desde sus PC de escritorio.
- Recuperación de datos más eficiente. Debido a que los datos almacenados se pueden respaldar rápidamente mediante la producción de copias, la recuperación de datos es más rápida y sencilla si un archivo se pierde o se daña.
- Reducción del espacio físico y mayor escalabilidad. Los archivadores físicos para compartir archivos, que ocupan mucho espacio con el tiempo, son innecesarios y aumentar la capacidad digital es sencillo.
- Potencialmente mayor protección de datos. Con las herramientas y funciones de seguridad avanzadas de la actualidad, existen muchas más opciones para salvaguardar y proteger datos particularmente sensibles de forma digital.
- La colaboración entre equipos se vuelve más sencilla. Los datos almacenados de forma centralizada son accesibles para todos los usuarios autorizados y pueden verse y compartirse entre equipos a medida que colaboran. - Mejora de la gestión de documentos. Los datos se pueden clasificar y organizar digitalmente con mayor facilidad, y esto se puede hacer desde una computadora de escritorio u otro dispositivo conectado.
- Mayor productividad y eficiencia del flujo de trabajo. Se necesita menos tiempo para guardar material digitalmente que imprimir páginas físicas y crear archivos que deben almacenarse en archivadores.

1.12.7 Soluciones emergentes de almacenamiento de datos

El campo del almacenamiento y la gestión de datos está en constante desarrollo. Los desarrollos más recientes en materia de almacenamiento en red pueden proporcionar soluciones integrales y con visión de futuro para las empresas que necesitan almacenar un volumen masivo de datos confidenciales.

Existen algunas alternativas de almacenamiento avanzadas para las empresas que necesitan un almacenamiento de big data más complejo.

Almacenamiento definido por software El almacenamiento de datos tradicional requiere el uso de hardware y software propietario. Cuando se necesita una mayor capacidad de almacenamiento, las empresas se apresuran a adquirir hardware adicional.

Por otro lado, el almacenamiento definido por software (SDS) desacopla la capa de software entre el lugar donde se guardan físicamente los datos y cómo se recuperan. Separar el software de almacenamiento de su hardware ayuda a aumentar la capacidad de almacenamiento en cualquier servidor estándar de la industria o sistema x86. Elimina el requisito de comprar más hardware propietario y emplea dispositivos de almacenamiento del mismo fabricante.

Al abstraer la capa de software, las organizaciones pueden colocar sus datos en cualquier lugar, con la capacidad de ampliar o reducir la capacidad según sea necesario. SDS ofrece beneficios adicionales, como administración automatizada, rentabilidad y conexión de varias fuentes de datos para crear una infraestructura de almacenamiento.

Virtualización de almacenamiento La virtualización de almacenamiento se refiere a la acumulación de capacidad de almacenamiento de varios dispositivos físicos y su posterior reasignación en un entorno virtualizado. Es la consolidación del almacenamiento físico de varios dispositivos en lo que parece ser un único dispositivo de almacenamiento controlado por una consola central.

Mediante el uso de software para localizar la capacidad de almacenamiento disponible, la tecnología agrega esa capacidad en un grupo de almacenamiento que las máquinas virtuales pueden usar en un entorno virtual.

La virtualización de almacenamiento, a diferencia de SDS, que separa la capa de software del hardware para establecer una infraestructura de almacenamiento, simplemente agrupa los recursos de almacenamiento para que aparezcan ante los usuarios como una única lectura o escritura normal en una unidad física.

Oculta la complejidad del sistema de almacenamiento, lo que permite a los usuarios y administradores realizar operaciones como copias de seguridad, archivado y recuperación de manera más eficiente y con menor tiempo. La virtualización de almacenamiento también puede ayudar a aumentar la capacidad de almacenamiento sin tener que comprar sistemas de almacenamiento adicionales.

Almacenamiento hiperconvergente El siguiente paso después de la virtualización del almacenamiento y el SDS es el almacenamiento hiperconvergente (HCS (hyper-converged storage)). El HCS utiliza la nube para integrar operaciones de computación, virtualización y almacenamiento en una unidad física que se puede administrar como un solo sistema.

Se trata de un almacenamiento definido por software porque cada nodo tiene una capa de software que ejecuta un software de virtualización idéntico al de todos los demás nodos del clúster. Este programa virtualiza y distribuye los recursos en cada nodo, lo que permite que el almacenamiento y otros recursos se utilicen como un solo grupo de almacenamiento o computación.

1.12.8 Otras tecnologías de almacenamiento emergentes

El futuro del almacenamiento de datos se está alejando de las unidades tradicionales por niveles y avanzando hacia los servicios combinados. Estos brindan a las empresas un mayor control sobre sus datos y minimizan la necesidad de grandes equipos de TI, ya que muchas actividades se pueden realizar de forma remota.

El **almacenamiento en la nube** al que los clientes pueden acceder desde muchos dispositivos es otro mercado en expansión que tiene el potencial de volverse aún más rápido y eficiente.

El **almacenamiento flash** y los chips dentro de las unidades SSD se están desarrollando como alternativas de almacenamiento confiables.

La **IA** también se está volviendo cada vez más común en las formas emergentes de almacenamiento de datos para manejar tareas repetitivas como mantener cronogramas de respaldo y establecer puntos de recuperación únicos para conjuntos de datos específicos.

1.12.9 Aprovechar los datos

Con la evolución de computadoras más rápidas, nuestra dependencia de los datos se ha multiplicado. Sin embargo, la pérdida de datos puede ocurrir en cualquier momento debido a varios factores, incluidos ransomware, fallas de hardware, cortes de energía, catástrofes naturales y errores humanos.

Para mantener el centro de datos y la red funcionando sin problemas y sin interrupciones durante el horario comercial habitual, las empresas deben tomarse el tiempo para realizar copias de seguridad de los datos y archivos críticos. El plan de recuperación de datos ante desastres de una empresa es como un seguro: uno espera no tener que usarlo nunca.

1.13 Data Lake vs Data Warehouse

En primer lugar: ¿qué tipos de datos existen (y por qué es importante)?

Para entender por qué elegiría un lago de datos en lugar de un almacén de datos, veamos qué tipos de datos existen y dónde puede almacenarlos:

Los datos **estructurados** (o “limpios”) son los datos clásicos de una hoja de cálculo: todo está bien y limpio (no quiere decir que no puede haber datos faltantes o un formato incorrecto) y los datos están organizados en una estructura similar a una tabla. Las bases de datos que almacenan este tipo de datos se denominan bases de datos relacionales: usamos SQL para administrar los datos en esas bases de datos. Algunos ejemplos de datos estructurados son los archivos .csv o Excel.

Los **datos semiestructurados**, como sugiere el nombre, incorporan algunos elementos de datos estructurados, aunque no están organizados en una estructura tabular. Sin embargo, contienen etiquetas y elementos para organizar los datos de una manera significativa y crear jerarquías. Las bases de datos que almacenan datos semiestructurados se denominan bases de datos no relacionales, como MongoDB.

Los **datos no estructurados**, que son la mayoría de los datos del mundo, son datos sin procesar que no siguen ningún esquema. Son los más ricos en información, pero deben limpiarse en la mayoría de los casos para que sean significativos. Algunos ejemplos de datos no estructurados son los archivos de video y audio, así como las fotos.

1.13.1 ¿Qué es un lago de datos?

Un lago de datos (Data Lake) es un repositorio que almacena todos los datos de su organización, tanto estructurados como no estructurados. Piense en él como un grupo de almacenamiento masivo para datos en su estado natural y sin procesar (como un lago). Una arquitectura de lago de datos puede manejar los enormes volúmenes de datos que la mayoría de las organizaciones producen sin la necesidad de estructurarlos primero. Los datos almacenados en un lago de datos se pueden usar para crear canalizaciones de datos para que estén disponibles para que las herramientas de análisis de datos encuentren información que sirva de base para decisiones comerciales clave.

1.13.2 Beneficios de Data Lake

Debido a que los grandes volúmenes de datos en un lago de datos no se estructuran antes de almacenarse, los científicos de datos capacitados o las herramientas de inteligencia empresarial de autoservicio de extremo a extremo pueden obtener acceso a una gama más amplia de datos mucho más rápido que en un almacén de datos.

- Se pueden almacenar volúmenes masivos de datos estructurados y no estructurados, como transacciones de ERP y registros de llamadas, de manera rentable.
- Los datos están disponibles para su uso mucho más rápido al mantenerlos en un estado sin procesar.
- Se puede analizar una gama más amplia de datos de nuevas formas para obtener información inesperada y que antes no estaba disponible.

Sus ingenieros de datos pueden crear canalizaciones de datos ETL y transformaciones de lectura de esquemas para que los datos almacenados en un lago de datos estén disponibles para análisis, ciencia de datos y aprendizaje automático. Las herramientas de creación de lagos de datos administrados lo ayudan a superar las limitaciones de los scripts lentos y codificados a mano y los escasos recursos de ingeniería.

En la actualidad, muchas empresas están adoptando Delta Lake, una capa de almacenamiento de código abierto que aprovecha la compatibilidad con ACID (atomicidad, consistencia, aislamiento y durabilidad) de las bases de datos transaccionales para mejorar la confiabilidad, el rendimiento y la flexibilidad en los lagos de datos. Es particularmente útil para escenarios que requieren capacidades transaccionales y aplicación de esquemas dentro de su lago de datos. Permite la creación de lagos de datos, que admiten tanto el almacenamiento de datos como el aprendizaje automático directamente en el lago de datos. Ofrece funciones como el manejo escalable de metadatos, el control de versiones de datos y la aplicación de esquemas para conjuntos de datos a gran escala, lo que garantiza la calidad y la confiabilidad de los datos para las tareas de análisis y ciencia de datos.

1.13.3 Definición de almacén de datos

Similar a un lago de datos, un almacén de datos es un repositorio de datos comerciales. Sin embargo, a diferencia de un lago de datos, solo los datos altamente estructurados y unificados residen en un almacén de datos para respaldar las necesidades específicas de inteligencia empresarial y análisis. Piense en ello como un almacén real, donde primero se procesan los contenidos, luego se organizan en secciones y en estantes (llamados data marts). Los datos de un almacén están listos para usarse para respaldar el análisis histórico y la generación de informes para informar la toma de decisiones en todas las líneas de negocios de una organización.

Un almacén de datos en la nube es una base de datos almacenada como un servicio administrado en una nube pública y optimizada para inteligencia empresarial y análisis escalables. Elimina la restricción de los centros de datos físicos y le permite aumentar o reducir rápidamente sus almacenes de datos para cumplir con los presupuestos y las necesidades comerciales cambiantes.

1.13.4 Beneficios del almacén de datos

Un almacén de datos ofrece enormes beneficios a las organizaciones, especialmente en lo que se relaciona con inteligencia empresarial y análisis. Después del trabajo inicial de limpieza y procesamiento, los datos almacenados en un almacén sirven como una "única fuente de verdad" consistente que es invaluable para el análisis de datos comerciales, la colaboración y una mejor comprensión. Tres ventajas principales de un almacén de datos incluyen:

- Se necesita poca o ninguna preparación de datos, lo que hace que sea mucho más fácil para los analistas y usuarios comerciales acceder a estos datos y analizarlos.
- Los datos precisos y completos están disponibles más rápidamente, por lo que las empresas pueden convertir la información en información más rápidamente.
- Los datos unificados y armonizados ofrecen una única fuente de verdad, lo que genera confianza en la información de los datos y la toma de decisiones en todas las líneas de negocio.

1.13.5 Data Lake vs Data Warehouse

En términos generales, puede imaginar un almacén de datos como una tabla de Excel y un lago de datos como un almacenamiento masivo como una memoria USB.

Aun así, muchas organizaciones usan tanto un lago de datos como un almacén de datos para cubrir el espectro de sus necesidades de almacenamiento de datos. Algunas optan por combinar las capacidades clave de cada uno mediante la implementación de un lago de datos. Echemos un vistazo en paralelo al lago de datos frente al almacén de datos, y cómo pueden trabajar en conjunto para brindar una solución integral de almacenamiento de datos para su empresa.

1.13.6 Data Lake vs Data Warehouse — 6 diferencias clave:

	Data Lake	Data Warehouse
1. Preparación de datos	Requieren poca o ninguna preparación, ya que aceptan cualquier tipo de datos sin restricciones. Esto también significa que para analizar los datos, primero debemos realizar un procesamiento considerable.	Los datos de entrada requieren más preparación para que coincidan con el esquema existente. Sin embargo, una vez que se han almacenado, los datos están disponibles para su análisis.
2. Flexibilidad	Proporciona alta flexibilidad y escalabilidad.	Un almacén de datos es menos flexible para nuevas fuentes de datos debido a su estructura rígida.
3. Mantenimiento	Requiere un alto mantenimiento y gestión de datos para evitar convertirse en un pantano de datos.	Requiere poco mantenimiento y alta solidez debido a su estructura.

	Data Lake	Data Warehouse
4. Estructura	Existe un esquema de lectura: la estructura se establece después de que se ingieren los datos. Por un lado, esto aumenta la flexibilidad, por otro lado, requiere más mantenimiento.	Existe el esquema de escritura: el esquema o la estructura se define antes de almacenar los datos. Esto aumenta los tiempos de procesamiento primero, pero una vez que se completa proporciona más solidez y consistencia.
5. Usuarios	Los lagos de datos son más difíciles de navegar, los datos son más complejos de procesar y el análisis no se puede implementar fácilmente. Los usuarios generalmente son científicos de datos, ingenieros de aprendizaje automático o analistas de datos.	Debido a su relativa simplicidad y estructura, los almacenes de datos se pueden administrar más fácilmente y los datos se pueden analizar más fácilmente. Los almacenes de datos son aptos para analistas comerciales e informes de KPI.
6. Tipos de datos	Un lago de datos contiene datos no estructurados. Esto significa que no hay restricciones sobre lo que se puede almacenar. Todo, desde imágenes, conjuntos de datos o archivos de audio, va. Esto significa que los usuarios tampoco están limitados por ninguna estructura impuesta de un lago de datos.	Un almacén de datos impone una estructura (columnas y filas) que se debe seguir. Como resultado, no podemos simplemente almacenar todos los datos que queremos. Los datos deben seguir las "reglas"/estructura del almacén de datos.

1.13.7 Formatos y esquemas de almacenamiento

Cuando se trabaja con big data, es fundamental comprender las diferencias entre los formatos de almacenamiento y cuándo aplicar esquemas. Exploremos cómo los almacenes de datos suelen utilizar "Schema-on-Write" y los lagos de datos emplean "Schema-on-Read".

Schema-on-Write

Schema-on-Write se refiere al proceso en el que se define el esquema de los datos antes de escribirlos en la base de datos. En un almacén de datos, los datos suelen estar estructurados y formateados en tablas similares a las bases de datos relacionales. Por ejemplo, los datos pueden almacenarse en formatos CSV o JSON, pero deben cumplir con el esquema definido, que incluye una estructura rígida de filas y columnas y tipos de datos.

Para un ejemplo práctico, insertar datos en una tabla SQL se vería así:

```
INSERT INTO sales (id, amount, quarter)
VALUES (1, 20000, 'Q1');
```

Aquí, el esquema está predefinido, lo que implica que cualquier dato que inserte debe ajustarse a esta estructura particular.

Schema-on-Read

Por el contrario, Schema-on-Read es un término que se utiliza predominantemente con lagos de datos. En estos entornos, los datos se almacenan en su forma original sin un esquema predefinido. Abarca una amplia gama de tipos de archivos, como CSV, JSON, ORC y Parquet. El esquema

se aplica más tarde, al momento de leer los datos, lo que permite una mayor flexibilidad y una variedad de tipos de datos.

Esto es particularmente útil cuando se trabaja con datos no relacionales que no encajan perfectamente en las tablas. El esquema se infiere o se aplica cuando se ejecuta una consulta para analizar los datos, y diferentes usuarios pueden aplicar diferentes esquemas según sea necesario para sus casos de uso específicos.

Por ejemplo, leer un archivo JSON con Schema-on-Read podría verse así:

```
df = spark.read.json("path/to/jsonfile.json")
df.printSchema()
```

En este caso, el marco de datos df tendría un esquema que se infiere del archivo JSON cuando se lee el archivo.

Ambos enfoques tienen sus propias ventajas: Schema-on-Write ofrece coherencia y control, mientras que Schema-on-Read proporciona flexibilidad y agilidad. Comprender cuándo utilizar cada método es fundamental para una gestión y un análisis de datos eficientes.

Creado por:

Isabel Maniega