Curso

Data Engineer

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Centro de e-Learning UTN.BA

Módulo 2:

Procesamiento avanzado y seguridad de datos

Unidad 3:

Seguridad de datos

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Presentación

La seguridad de datos en el contexto de la Ingeniería de Datos es un aspecto fundamental para proteger la información de posibles amenazas y riesgos. Como Data Engineer, es nuestra responsabilidad garantizar que los datos estén seguros y protegidos durante su almacenamiento, procesamiento y transmisión.

La seguridad de datos se refiere a las medidas y prácticas que se implementan para mantener la confidencialidad, integridad y disponibilidad de la información.

Existen diferentes aspectos de seguridad de datos que debemos considerar. Por ejemplo, la encriptación es una técnica que convierte los datos en un formato ilegible para cualquier persona que no tenga la clave adecuada. Es como tener una caja fuerte con una combinación secreta: solo las personas autorizadas pueden descifrar los datos.

Otro aspecto importante es el control de acceso. Aquí es donde definimos quién tiene permiso para acceder a los datos y en qué medida. Es como tener una puerta con una cerradura y solo permitir que las personas autorizadas tengan la llave correcta. Esto nos ayuda a evitar que personas no autorizadas accedan a los datos y mantengan su confidencialidad.

Por último, la anonimización de datos es otra práctica relevante. Implica la transformación de los datos de manera que no sea posible identificar a los individuos a los que pertenecen. Esto se logra mediante la eliminación o alteración de información personal identificable (PII) como nombres, direcciones, números de teléfono, etc.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bloques temáticos

1. Encriptación
2. Anonimización
3. Controles de acceso

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Encriptación

La encriptación es una técnica esencial en la seguridad de datos que nos ayuda a proteger la confidencialidad de la información. Básicamente, consiste en convertir los datos en un formato ilegible llamado "texto cifrado" mediante el uso de algoritmos matemáticos. Para poder leer y comprender el texto cifrado, se necesita una clave especial conocida como "clave de encriptación" que solo la persona autorizada posee.

La encriptación es una herramienta poderosa para proteger la confidencialidad de los datos, tanto en reposo (almacenamiento) como en tránsito (transmisión). Se utiliza ampliamente en aplicaciones y servicios que manejan información sensible, como sistemas de pago en línea, comunicaciones seguras y almacenamiento en la nube. Sin embargo, es importante destacar que la encriptación no es una solución completa de seguridad por sí sola. También se deben implementar otras medidas de seguridad, como el control de acceso adecuado y la gestión segura de las claves de encriptación, para garantizar una protección sólida de los datos.

Como Data Engineer, nuestra responsabilidad, en términos de encriptación, se centra en **asegurar que los datos estén protegidos durante su almacenamiento, procesamiento y transmisión**. Algunos ejemplos de cómo un/a Data Engineer puede utilizar la encriptación son:

- Encriptación de datos en reposo (at-rest): Cuando los datos se almacenan en una base de datos o en sistemas de almacenamiento, hay que asegurarse de implementar encriptación para protegerlos de accesos no autorizados. Esto implica utilizar técnicas de encriptación para transformar los datos en un formato ilegible mientras están en reposo. Por ejemplo, se puede utilizar algoritmos de encriptación para cifrar los archivos o bases de datos, asegurando que solo las personas autorizadas con la clave de encriptación puedan acceder y leer los datos.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Existen tecnologías que ofrecen la encriptación en reposo de forma interna. Muchas soluciones de almacenamiento en la nube brindan funciones incorporadas de encriptación en reposo para proteger los datos almacenados. Estos sistemas pueden utilizar métodos de encriptación automáticos y transparentes que cifran los datos en el nivel del almacenamiento subyacente, sin requerir una intervención directa del Data Engineer.

- Encriptación de datos en tránsito: Durante la transmisión de datos a través de redes, es fundamental garantizar que la información no sea interceptada o manipulada por terceros. Para lograr esto, puedo utilizar protocolos seguros, como HTTPS, que utilizan encriptación para proteger las comunicaciones entre los diferentes sistemas. Esto implica asegurarse de que los datos estén cifrados mientras se transmiten y solo puedan ser desencriptados por el destinatario legítimo.
- Gestión de claves de encriptación: La seguridad de la encriptación depende en gran medida de la gestión adecuada de las claves de encriptación. Como Data Engineer, debemos implementar prácticas seguras para generar, almacenar y administrar las claves de encriptación. Esto incluye utilizar técnicas de gestión de claves, como el uso de claves fuertes y aleatorias, rotación regular de claves y almacenamiento seguro de claves en bóvedas.

  Azure Key Vault es una herramienta adecuada y muy útil para un/a Data Engineer en la gestión de claves. Es un servicio de administración de claves y secretos en la nube proporcionado por Microsoft Azure. Entre sus características se encuentran:

  - almacenamiento seguro de claves

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

- generación y rotación de claves: puede generar claves fuertes y aleatorias y permite la rotación periódica de claves para mejorar la seguridad y cumplir con las mejores prácticas
- control de acceso granular para definir y gestionar quién tiene acceso a las claves almacenadas.

Anonimización

La anonimización de datos es un proceso que se utiliza para proteger la privacidad de los individuos al eliminar o transformar la información personal identificable (conocida como PII, por sus siglas en inglés) de los conjuntos de datos, de manera que los datos resultantes no puedan ser utilizados para identificar a personas específicas. El objetivo principal de la anonimización es preservar la utilidad y la integridad de los datos, al tiempo que se minimiza el riesgo de revelación de la identidad de los individuos.

Existen varias técnicas de anonimización de datos que se pueden utilizar, algunas de ellas son:

- Supresión: Consiste en eliminar completamente los datos que podrían identificar a una persona. Esto implica eliminar campos o columnas como nombres, direcciones, números de identificación o cualquier otro dato personal que pueda relacionarse directamente con un individuo.
- Masking: Consiste en ocultar de forma parcial o total ciertos caracteres de información sensible, como números de tarjetas de crédito o direcciones de correo electrónico, con el objetivo de preservar la confidencialidad de los datos mientras se mantiene la estructura y la utilidad general de la información. En el caso de tarjetas de crédito, por ejemplo, se pueden reemplazar algunos dígitos con caracteres genéricos, como "X" o "*", manteniendo solo los primeros y/o últimos dígitos visibles. Esto ayuda a ocultar parte de la información sensible sin perder completamente la estructura y el formato del número de tarjeta.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

- Generalización: Implica reemplazar valores específicos por categorías más amplias o rangos generales. Ayuda a proteger la privacidad de los individuos al reducir el nivel de detalle de la información sin perder la utilidad general de los datos. Sin embargo, se deben considerar cuidadosamente los riesgos y los posibles impactos en la reidentificación antes de aplicar la generalización a los datos sensibles.
- Uso de datos sintéticos o ficticios generados artificialmente, que mantienen ciertas características y distribuciones estadísticas similares a los datos originales, pero no están asociados directamente con personas reales.

Controles de acceso

Como Data Engineer, una responsabilidad clave es asegurar un adecuado control de acceso a los datos. El control de acceso se refiere a la gestión de quién tiene permiso para acceder y manipular los datos en un entorno como un Data Lake o un Data Warehouse, por ejemplo.

Imaginemos una casa con varias habitaciones. Cada habitación contiene información valiosa y privada. Como Data Engineer, debemos asegurarnos de que solo las personas autorizadas puedan acceder a cada habitación y ver su contenido. Esto se logra mediante la implementación de controles y mecanismos de seguridad.

Para garantizar que las personas autorizadas puedan acceder y manipular los datos, hay que realizar varias tareas como las siguientes:

- Identificación de usuarios y roles: Relevar quiénes son los usuarios que necesitan acceder a los datos. Los roles determinan qué acciones pueden realizar los usuarios y qué datos pueden ver.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

- Autenticación: Definir mecanismos para verificar la identidad de los usuarios antes de permitirles el acceso a los datos. Esto puede incluir el uso de contraseñas, claves de acceso, etc.
- Autorización: Una vez que un usuario ha sido autenticado, es importante asegurarse de que solo tenga acceso a los datos y las acciones que le corresponden. Se implementan medidas para definir y controlar los permisos de cada usuario, asegurando que solo puedan realizar las operaciones permitidas.

Autorización

En sistema como un data lake, donde se almacenan diversos tipos de datos en su forma original, el control de acceso se puede lograr mediante el uso de listas de control de acceso (ACL, por sus siglas en inglés) o el control de acceso basado en roles (RBAC, por sus siglas en inglés).

- ACL (Listas de Control de Acceso): Con ACL se definen permisos de acceso específicos para cada objeto o recurso en el data lake. Esto implica asignar permisos individuales a usuarios o grupos para leer, escribir o ejecutar acciones en archivos o carpetas específicas. Por ejemplo, podemos permitir que un grupo de analistas de datos tenga acceso de lectura a un conjunto de datos, mientras que solo algunos ingenieros tengan permiso de escritura en ese mismo conjunto de datos.
- RBAC (Control de Acceso Basado en Roles): Con RBAC, en lugar de asignar permisos individuales a usuarios, se crean roles y se asignan permisos a esos roles. Luego, a los usuarios se les asigna los roles apropiados. Esto simplifica la administración de permisos al agrupar a los usuarios en roles que se ajusten a sus responsabilidades. Por ejemplo, podemos crear un rol "Analista de Datos" con permisos de lectura y un rol "Ingeniero de Datos" con permisos de escritura, y luego asignar a los usuarios a los roles correspondientes.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Conclusión

En conjunto, la encriptación, la anonimización y el control de acceso son pilares fundamentales para garantizar la seguridad de los datos en la Ingeniería de Datos. Al implementar estas prácticas, podemos salvaguardar la integridad, confidencialidad y disponibilidad de la información, protegiendo la privacidad de las personas y promoviendo un entorno seguro y confiable para la gestión de datos.

Es importante tener en cuenta que la seguridad de datos es un campo en constante evolución, y como Data Engineers, debemos estar actualizados sobre las mejores prácticas y las tecnologías emergentes para enfrentar los desafíos y riesgos en este ámbito en constante cambio.

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning

Bibliografía utilizada y sugerida

- Data Security: Definition, Importance, and Types | Fortinet. (s. f.). Fortinet. https://www.fortinet.com/resources/cyberglossary/data-security
- Role-based access control (RBAC). (2023, 23 marzo). TechTarget. https://www.techtarget.com/searchsecurity/definition/role-based-access-control-RBAC

CENTRO DE E-LEARNING UTN BA
Medrano 951 CABA, Buenos Aires Argentina
(1179) // tel +54 11 7078 – 8073 / fax +54 11 4032 0148
www.sceu.frba.utn.edu.ar/e-learning