aws
COMMUNITY DAY
CHILE

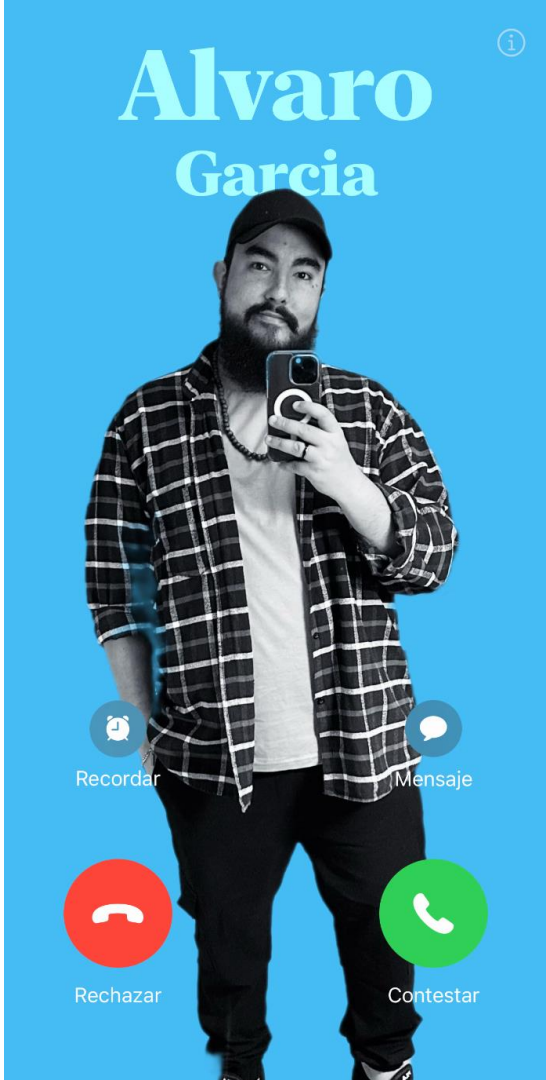
Navegando en aguas tranquilas :

Cómo evitar que tu Data Lake se convierta en un Data Swamp

Alvaro Garcia

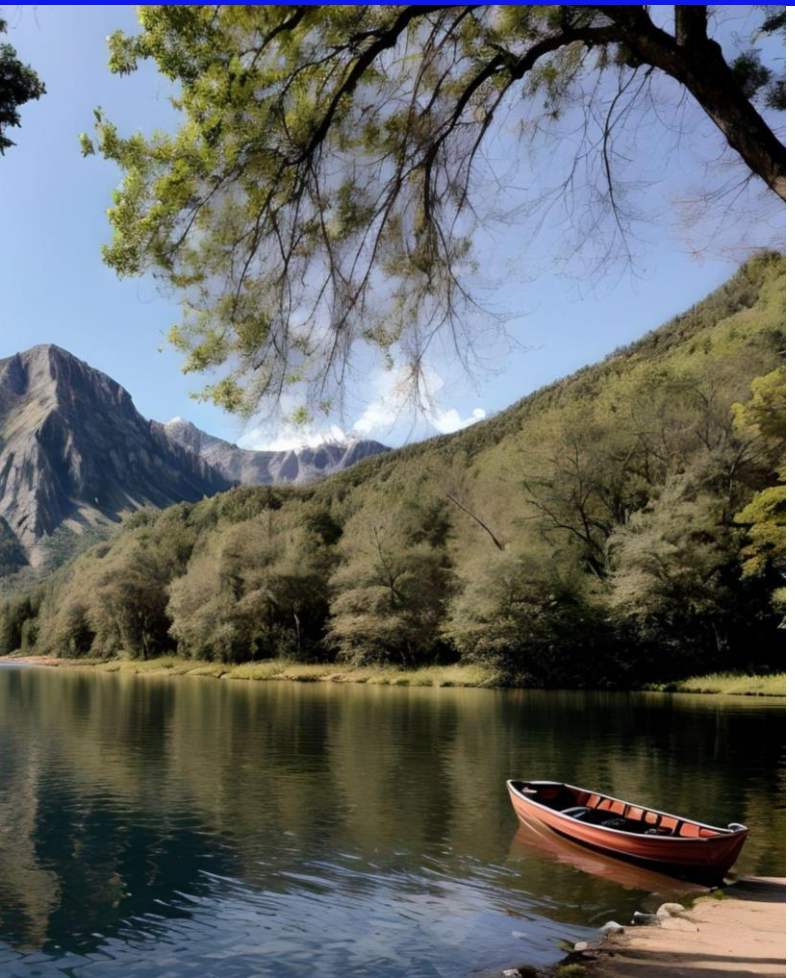
TL Data at Cloudhesive

- 13 Años en consultoría de datos.
- Entusiasta de la seguridad informática.
- Uruguayo.



AGENDA

- Introducción
- Comprendiendo los conceptos clave
- Evitar el desorden
- Aumentar la calidad
- Mantener la seguridad
- Potenciar el interés y la adopción
- Conclusión
- Q&A



Dato = Activo

Importancia del Data Lake

Hoy en día, los datos son el activo más valioso de una organización, y un Data Lake es la herramienta que permite almacenar y analizar eficazmente grandes cantidades de datos sin procesar en su formato original, marcando la diferencia entre el éxito y el estancamiento.

Diferencias entre Data Lake

y dolores de cabeza.

Diferencias entre Data Lake y Data Swamp

Data Lake

- Organizado y estructurado.
- Alta calidad datos.
- Seguro y protegido.
- Amplia adopción empresarial.
- Valioso para negocios.

Data Swamp

- Desorden y confusión.
- Baja calidad datos.
- Riesgo seguridad datos.
- Baja adopción empresarial.
- Valor limitado datos.

¿Cómo está planteada esta charla?

Situación

Vamos a plantear una situación específica en la que indagaremos en las causas y problemas asociados a no prestarle atención.

Estrategia

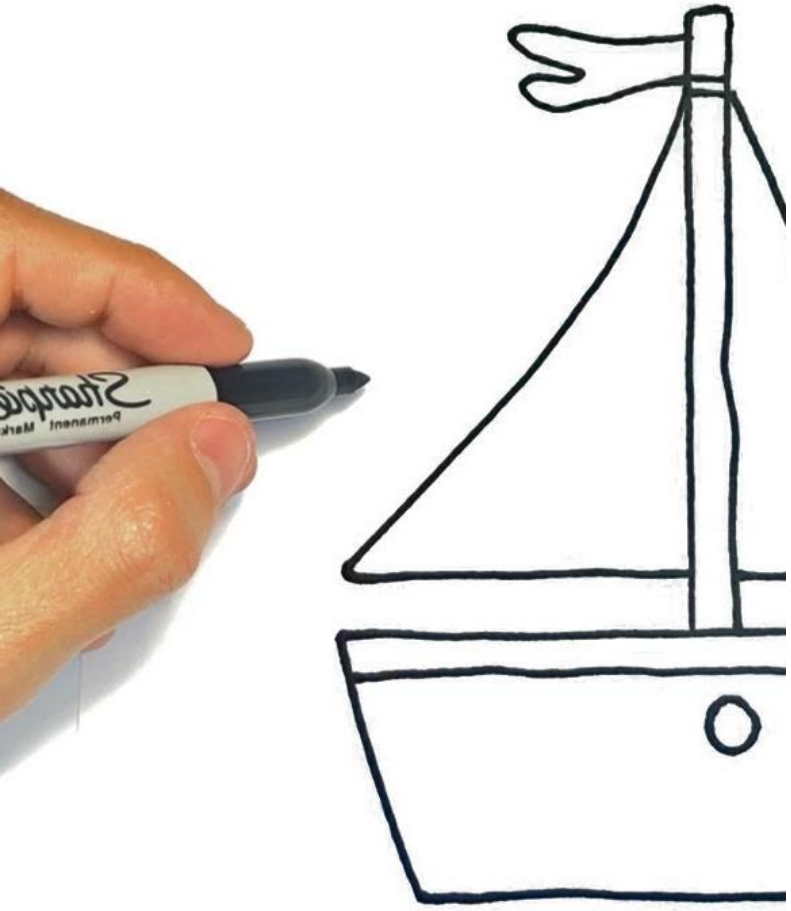
Plantearemos a nivel estratégico cual es la forma más eficiente de afrontar esa situación.

Servicios AWS

Repasaremos dentro de los servicios de aws cuales nos ofrecen herramientas necesarias para afrontar estos cambios.

Evitar el Desorden

Manteniendo la Claridad en Tu Data Lake



Simplicidad

Y enfoque en el negocio

La simplicidad en la gestión de datos es la clave para un Data Lake organizado. Al centrarse en las necesidades del negocio y eliminar el exceso de complejidad, se crea un entorno propicio para la eficiencia y la toma de decisiones informadas.

Situación 1

La Importancia de la Organización

- Gran acumulación de datos.
- Datos en dispersos o duplicados
- Desperdicio de recursos
- Dificultad en la búsqueda

Situación 1 : Estrategia

01

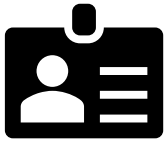
Estructura Lógica

Crea una jerarquía de carpetas clara y coherente para clasificar los datos según categorías lógicas, como tipo, departamento o proyecto.

02

Utilizar Metadatos

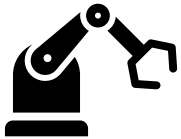
Aplica metadatos relevantes, como fecha, propietario y descripción, a cada conjunto de datos para facilitar su búsqueda y comprensión.



03

Automatizar

Automatizar la catalogación y organización de datos en función de reglas predefinidas y metadatos, mejorando la eficiencia de la gestión de datos en tu Data Lake.



Situación 1 : Servicios AWS

S3

- Principal servicio de almacenamiento para un data lake.
- Serverless.
- Escalable.
- Costo eficiente.

Glue Catalog

- Catálogo Centralizado
- Organización de Metadatos
- Soporte para Múltiples Fuentes
- Integración con Servicios de AWS

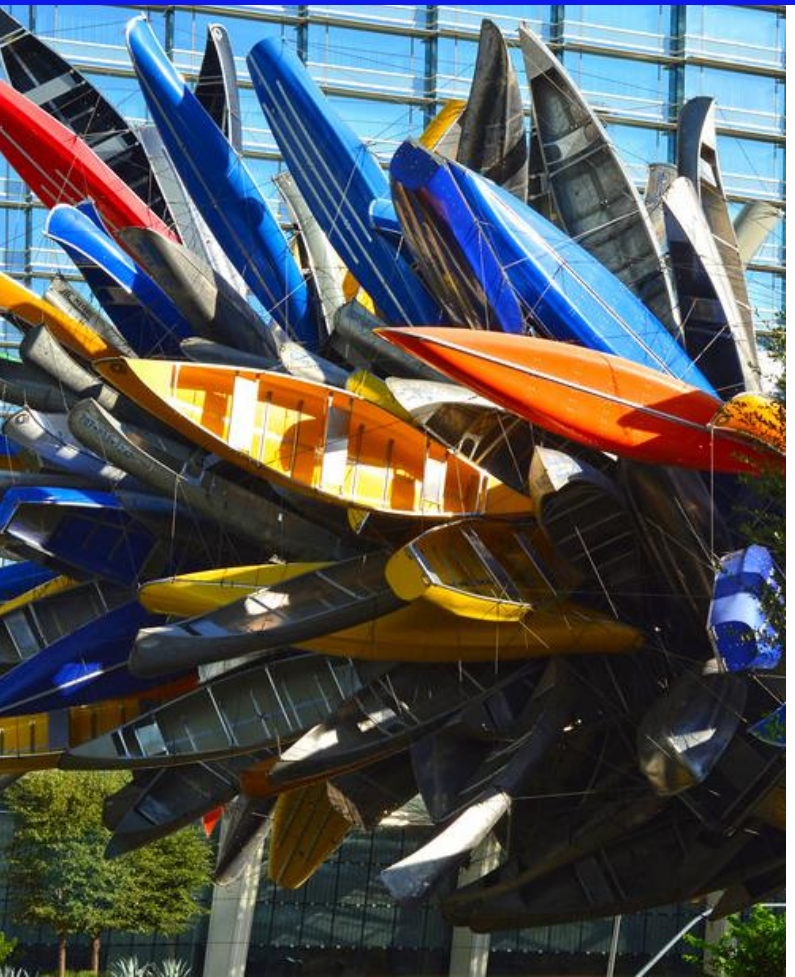
Lambda + Glue Jobs

- Procesamiento paralelo
- Autoescalable
- Serverless
- Event Driven
- Facilmente Integrables

Aumentar la calidad datos

Manteniendo la confiabilidad del Data Lake

Aumentar la calidad datos



Calidad

Sobre cantidad

En la gestión de un Data Lake, es fundamental recordar que la calidad de los datos es prioritaria sobre la cantidad. Más datos no siempre significan mejores decisiones, pero datos de alta calidad impulsan análisis precisos y resultados confiables, proporcionando un valor real a la organización.

Situación 2

Calidad Sobre cantidad

- Datos incorrectos.
- Problemas de integridad.
- Dificultad en la búsqueda.
- Información incompleta.

Situación 2 : Estrategia

01

Evaluar la calidad

Realiza una evaluación exhaustiva de la calidad de los datos existentes para identificar problemas, como datos duplicados, incoherencias o falta de integridad.

02

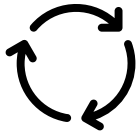
Limpieza y Normalización

Implementar procesos de limpieza y normalización de datos basados en reglas de negocio pre configuradas.

03

Validaciones Continuas

Establecer validaciones y reglas automatizadas para garantizar la calidad de los datos en tiempo real.



Situación 2 : Servicios AWS

Glue DataBrew

- Perfil de datos.
- Limpieza fácil.
- Visualización interactiva.
- Sin código.
- Serverless

Glue Jobs

- Procesamiento paralelo
- Autoescalable
- Serverless
- Event Driven
- Facilmente Integrables

Lambda

- Event Driven
- Autoescalable
- Serverless
- Eventos en tiempo real.
- Código personalizado.

Mantener la seguridad

Un imperativo en la era digital.

Gobierno

Vs Democratización

Anteriormente se manejaba el concepto de que la democratización iba en contra del gobierno de los datos, hoy en día utilizando tecnologías disponibles en AWS podemos lograr que aumentar el acceso a los datos por parte de los usuarios, no suponga un riesgo de governanza para la compañía.

Situación 3

La Importancia de la Seguridad

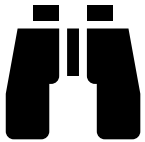
- Gobierno de datos Inexistente.
- Información sin protección.
- Datos sensibles en riesgo.
- Información incompleta.

Situación 3 : Estrategia

01

Evaluar Riesgos y compliance

Realizar una evaluación exhaustiva de los riesgos de seguridad y los requisitos de cumplimiento que afectan a tus datos.



02

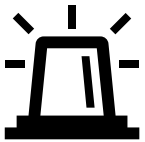
Políticas de control y acceso

Definir y aplicar políticas de acceso y controles de seguridad adecuados para proteger los datos y restringir el acceso no autorizado.

03

Monitoreo y detección continua

Establecer sistemas de monitoreo y detección de amenazas en tiempo real para identificar y responder rápidamente a cualquier actividad sospechosa.



Situación 3 : Servicios AWS – Evaluación de riesgos

Security Hub

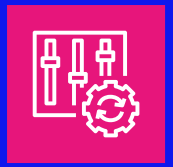
- Centraliza seguridad.
- Evaluación cumplimiento normativo.
- Alertas de seguridad.
- Identifica vulnerabilidades.
- Monitoriza en tiempo real.

Amazon Macie

- Privacidad datos.
- Aprendizaje automático.
- Protege información confidencial.
- Descubrimiento automático.
- Cumplimiento normativo

AWS Config

- Evalúa configuraciones
- Cumplimiento políticas.
- Audita cambios
- Identifica desviaciones.
- Rastreo y registros.



Situación 3 : Servicios AWS – Control de Accesos

IAM

- Control de acceso.
- Políticas granulares.
- Gestión de identidades.
- Seguridad de recursos.
- Restricción de permisos.

Lake Formation

- Gestión de Data Lake.
- Control de acceso.
- Cumplimiento normativo.
- Definición de políticas.
- Facilita seguridad.
- Amigable al usuario.

Situación 3 : Servicios AWS – Monitoreo y detección

GuardDuty

- Detección de amenazas.
- Monitoreo continuo.
- Análisis de comportamiento.
- Alertas de seguridad.
- Protege el Data Lake.

CloudWatch

- Monitoreo de recursos.
- Alarmas personalizadas.
- Registros y métricas.
- Alertas en tiempo real.
- Seguimiento de actividades.

Potenciar el interés y la adopción

Foco en la adopción del data lake en todos los niveles.

Situación 4

La Importancia de la UI/UX y el MKT

- Silos de información.
- Baja adopción del Data Lake.
- Necesidad de compromiso y uso.
- Excel

Situación 4 : Estrategia

01

Comprensión de Necesidades

Realiza encuestas y entrevistas para comprender las necesidades y desafíos específicos de los usuarios en relación con los datos.

02

Personalización y Facilitación

Personaliza las interfaces de acceso al Data Lake para que sean intuitivas y fáciles de usar, y proporciona acceso rápido a los datos relevantes.

03

Fomento del Uso y la Colaboración

Fomenta la colaboración y el intercambio de conocimientos entre los usuarios, promoviendo el uso activo del Data Lake para resolver problemas y tomar decisiones informadas.

Situación 3 : Servicios AWS

Amazon Pinpoint

- Comunicación de clientes.
- Armado de formularios.
- Campañas de marketing.
- Analítica de interacciones.
- Personalización de mensajes.

QuickSight

- Visualización de datos.
- Paneles interactivos.
- Informes en tiempo real.
- Integrable a webs propias.
- Detección de anomalías.
- Narrativas Automáticas

AWS Data Zone

- Integra Governanza y seguridad
- Diseñada para la colaboración.
- Busqueda y Consulta
- Portal fuera de la consola.

Conclusión

Recapitulemos los puntos clave.

Recapitulemos los puntos clave

1. **Organización y Clasificación:** Aprender la importancia de etiquetar, categorizar y catalogar tus datos para mantener el orden en tu Data Lake.
2. **Seguridad en los Datos:** Comprender cómo la encriptación, la autenticación y la gestión de identidades son esenciales para proteger tu Data Lake contra amenazas.
3. **Uso y Colaboración:** Fomentar la colaboración y el uso activo de tu Data Lake, involucrando a equipos de toda la organización.
4. **Valor de la Herramienta:** Demostrar el valor de tu Data Lake mediante la creación de informes impactantes y visualizaciones que impulsen la toma de decisiones informadas.
5. **Capacitación y Soporte:** Proporciona capacitación continua y soporte sólido para garantizar que tu equipo pueda utilizar eficazmente el Data Lake.

Estos son los puntos clave que debes recordar para mantener tu Data Lake en buen estado y evitar que se convierta en un Data Swamp.

aws
COMMUNITY DAY
CHILE

¡Muchas gracias!

SE INICIA LA RONDA DE PREGUNTAS.