#DATA2023 Workshop – Delta Lakes

## Reglas básicas

- Tener previamente acceso a AWS con acceso programático
- Tener conocimiento básico de Python
- Haber descargado el repo del taller <u>aquí</u>
- Cada participante debe generar su propio ambiente







### **Agenda**

- 1. Delta Lakes 101
- 2. Delta lakes en AWS
- 3. Desplegando recursos en AWS (Terraform)
- 4. S3 + Glue para Delta Tables
- 5. Glue in Deep (EQTL)
- **6.** Athena en Delta Lakes
- 7. Analítica con Delta Lakes
- 8. Orquestación con Airflow\*



# **Delta Lakes 101**



### ¿Que son los Delta Lakes?

**Delta Lake** es una capa de almacenamiento de código abierto creada por Databricks que proporciona capacidades avanzadas y mejoras en los **Data Lakes** tradicionales. Está diseñado para abordar los desafíos asociados con el procesamiento de grandes volúmenes, la confiabilidad y la gestión del ciclo de vida de los datos.



### ¿Por qué usar Delta Lakes?

### Beneficios

- Confiabilidad e Integridad de los datos: Al realizar transacciones ACID, garantiza la confiabilidad e integridad de los datos
- Escalabilidad: Están diseñados para manejar grandes volúmenes de datos. Utilizan formatos de archivos optimizados (Parquet) y técnicas de indexación que permiten realizar consultas eficientes
- Time Travel y Versionado: Permite realizar consultas en diferentes momentos del tiempo, rastrear cambios y comparar diferentes versiones de la tabla. Esta función es principalmente para auditorías y compliance
- Schema Evolution: Permite los cambios en la estructura con la que se almacenan los datos, lo que permite a las tablas evolucionar sin que se pierda la compatibilidad con los datos existentes

- Consistencia y Control de Concurrencia: Al utilizar MVCC, garantiza que las lecturas y escrituras concurrentes no interfieran entre sí
- Unificación de Batch y Stream processing: Permite que un mismo pipeline maneje datos por lotes o en tiempo real, unificando así el procesamiento y reduciendo la necesidad de solucione específicas y los esfuerzos de desarrollo
- Compatibilidad e integración: Esta diseñado para ser totalmente compatible con Data Lakes, y funge como una evolución de los mismos, donde también se integra fácilmente con herramientas como Spark, Hadoop y Hive



### ¿Problemas para usar Delta Lakes?

### Desventajas

- Curva de Aprendizaje: El uso de Delta Lakes implica familiarizarse con nuevos conceptos y características
- Costo de Rendimiento: Aunque Delta Lakes ofrecen muchos beneficios, algunas operaciones, como actualizaciones o eliminaciones frecuentes de pequeñas cantidades de datos, pueden tener un impacto en el rendimiento
- Complejidad: Delta Lakes introducen complejidad adicional en la arquitectura de datos
- Uso Adicional de Almacenamiento: Delta Lakes almacenan metadatos y registros de transacciones junto con los datos reales, lo que puede resultar en un mayor uso de almacenamiento en comparación con los formatos de archivo tradicionales

- Dependencia de un Proveedor: Aunque Delta Lake es un formato abierto, su adopción puede generar cierta dependencia de un proveedor, especialmente si se utilizan características o integraciones propietarias específicas de Delta Lake
- Compatibilidad y Soporte de Ecosistema: Aunque Delta Lake ha ganado popularidad, no todas las herramientas o frameworks de procesamiento de datos pueden ser compatibles de manera nativa con Delta Lake
- Madurez y Soporte de la Comunidad: Delta Lake es una tecnología relativamente nueva en comparación con soluciones de almacenamiento de datos más establecidas



### ¿Como son almacenados los datos?

### Evolución de los Data Lakes

### Parquet

- Almacenamiento columnar
- Ofrece un mejor compresión para el almacenamiento (GZIP, Snappy)
- Codificación de datos avanzada (RLE)
- Filtrado "pushdown" (Se hacen las consultas en los archivos in cargar todo en memoria)
- Estructura de metadata
- Compatible con múltiples paltaformas (Spark, Hadoop, Hive)

### Log Files

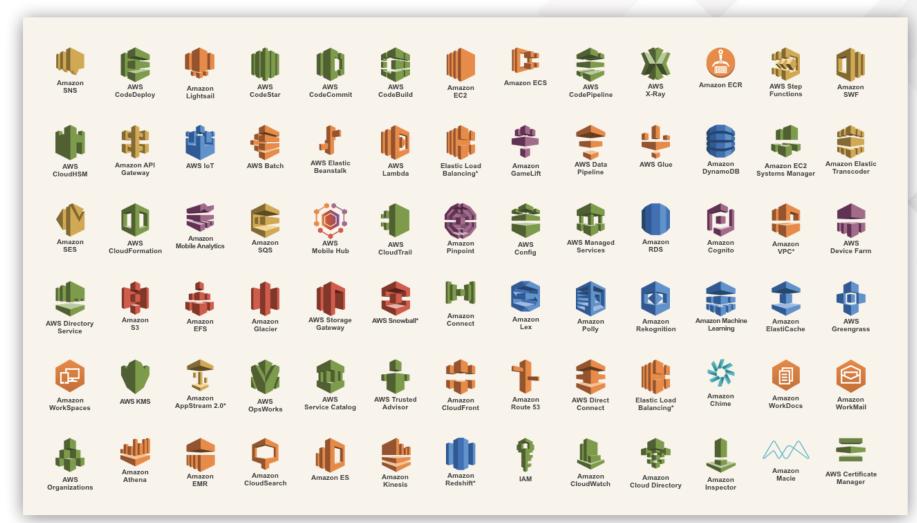
- Registro de cambios (Insert, Update, Delete)
- Atomicidad y consistencia
- Time travel (Indicador de estado de tablas en el tiempo)
- Reprocesamiento y recuperación
- Gestión de metadata (Esquema, versiones de tabla, etc)
- Control de concurrencia



# Delta Lakes en AWS



### ¿Qué usar en un mar de posibilidades?





### Enfocados en lo básico

Las aplicaciones básicas para Data Lakes, son útiles para la evolución a Delta Lakes



- Almacenamiento de objetos escalable
- Alta disponibilidad
- Variedad de opciones de seguridad
- Costo controlado en base a los datos en uso
- Capacidad de "Cold Storage"



- Serverless computing
- Ejecución a demanda en tiempo real o batch
- Alta compatibilidad con diferentes lenguajes
- Alta disponibilidad
- Fácil monitoreo de ejecuciones



- Serverless ETL
- Data catalog centralizado
- Data discovery
- Fácil monitoreo de ejecuciones
- Framework nativo de calidad de datos

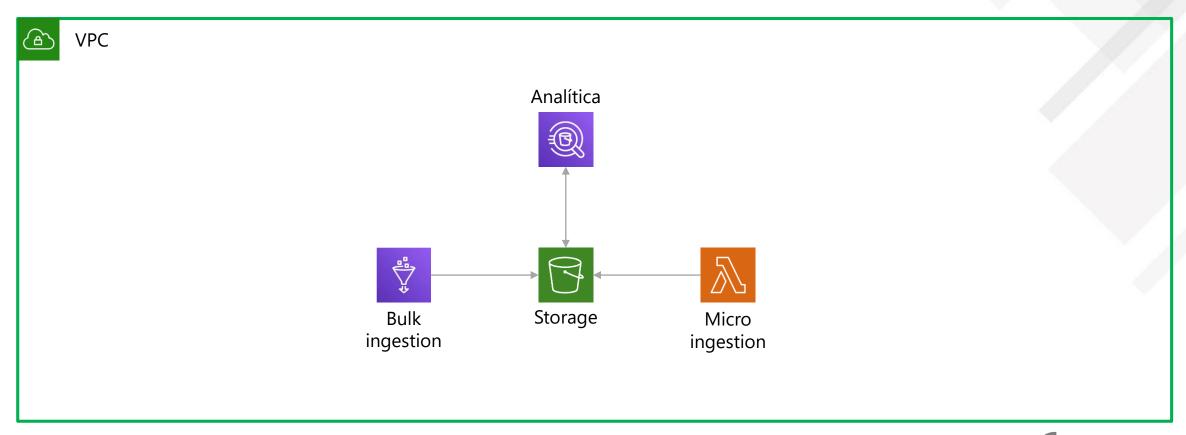


- Autoescalamiento automático
- Integración nativa con S3
- Procesa datos estructurados y no estructurados
- Fácil seguridad y control de acceso
- Monitoreo de transacciones históricas



### Basados en una arquitectura simple pero escalable

Serverless, permite disponer de mayores recursos sin necesidad de definirlos de entrada





# DEMO