Módulo Profesional: Big Data Aplicado

Change Data Capture
Debezium

ÍNDICE

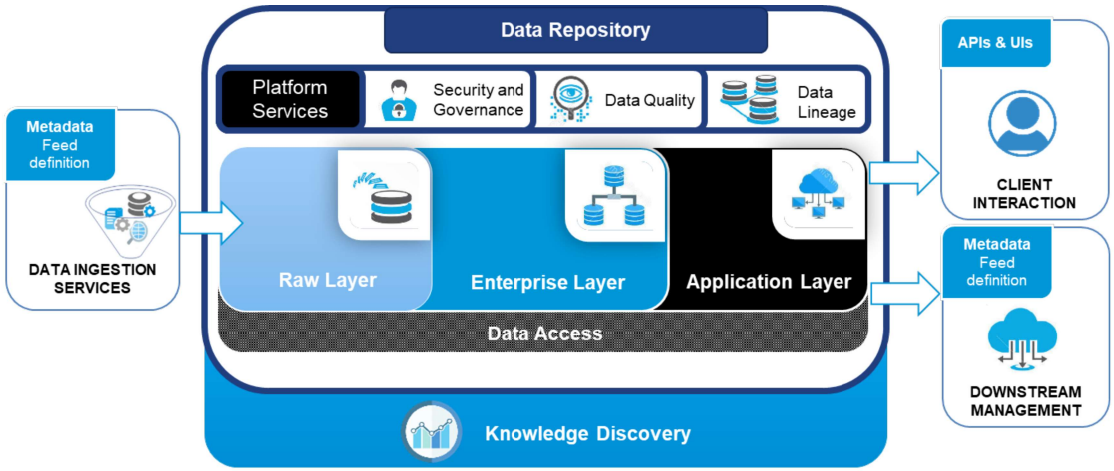
1. Introducción
- Qué es Change Data Capture (CDC)
- Herramientas disponibles
- Casos de uso

2. Debezium
- Introducción
- Características
- Configuración del entorno
- Descarga de Base de Datos en tiempo real

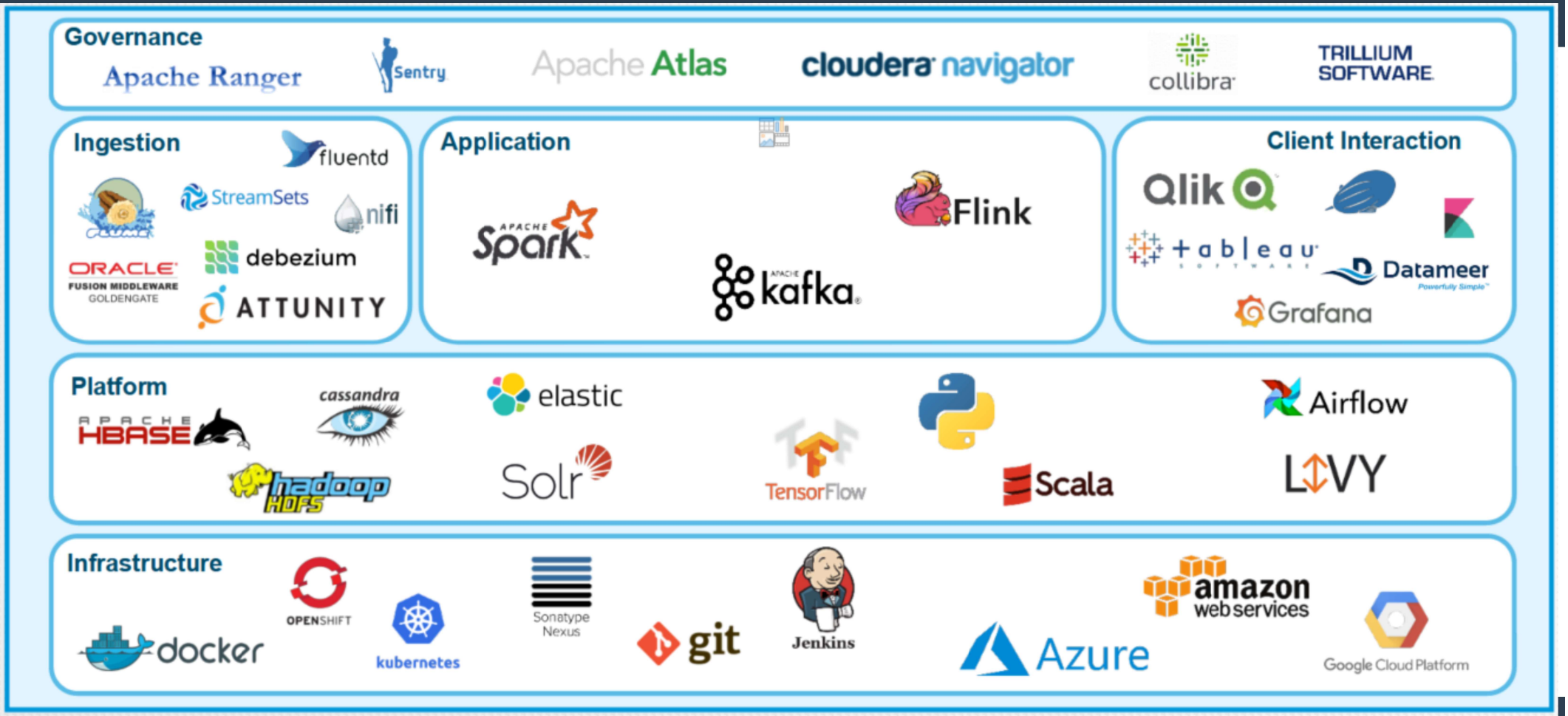
3. Ejercicios

INTRODUCCIÓN

La ingestión de datos es el proceso de adquirir datos de sistemas de origen y enviarlos (y/o almacenarlos) en un sistema de destino.
Una correcta Ingesta de Datos es clave para obtener mejores y más rápidos insights, reduciendo nuestro time-to-market.



TECHNOLOGY LANDSCAPE (PANORAMA TECNOLÓGICO)

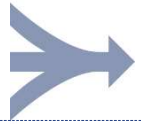


¿QUÉ ES CDC?

- CDC son las siglas en inglés de Captura Cambios de Datos
- Es un mecanismo eficaz para extraer datos de las fuentes
- CDC se refiere al proceso para identificar y capturar los cambios realizados en una base de datos
- Algunos cambios pueden propagarse a otro repositorio de datos (por ejemplo, Data Lake)

Automatizado

Eficiente
Extracción de datos
(no intrusiva)



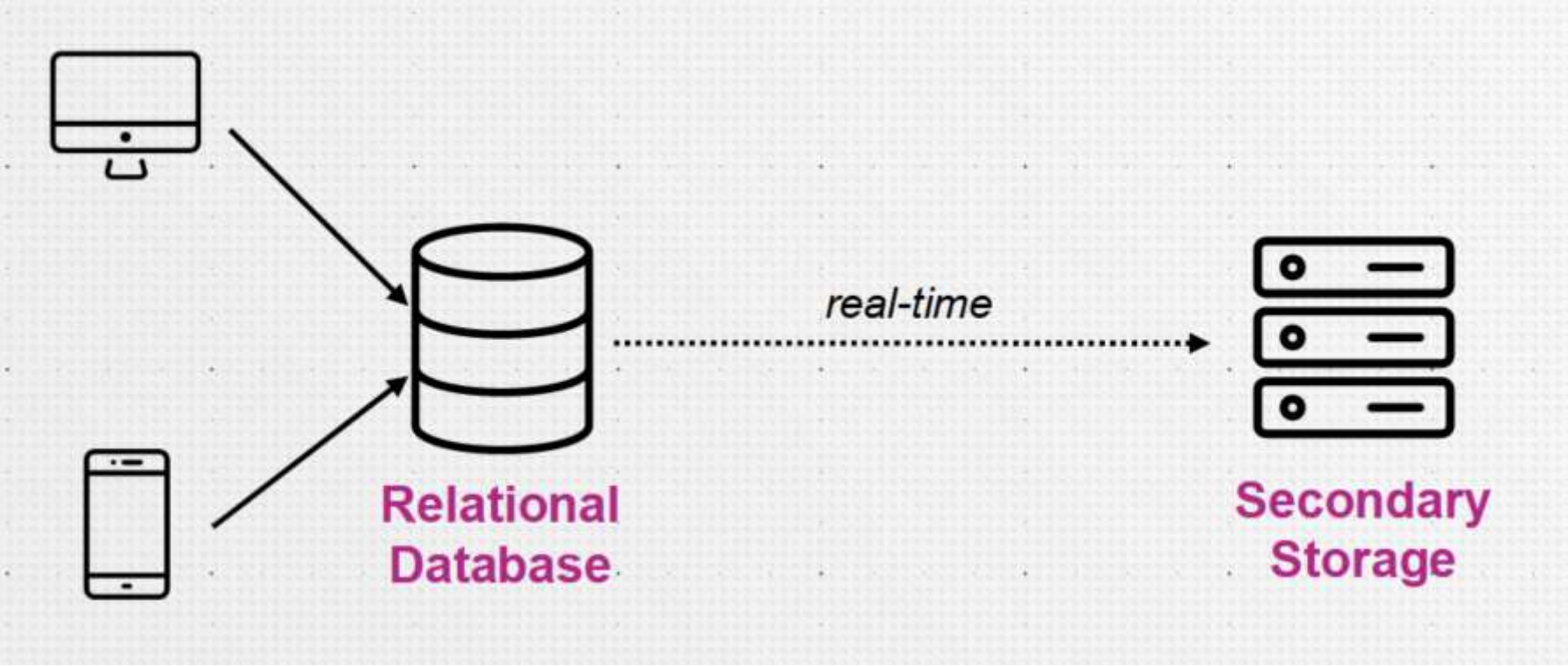
Streaming en tiempo real

Integración con múltiples fuentes

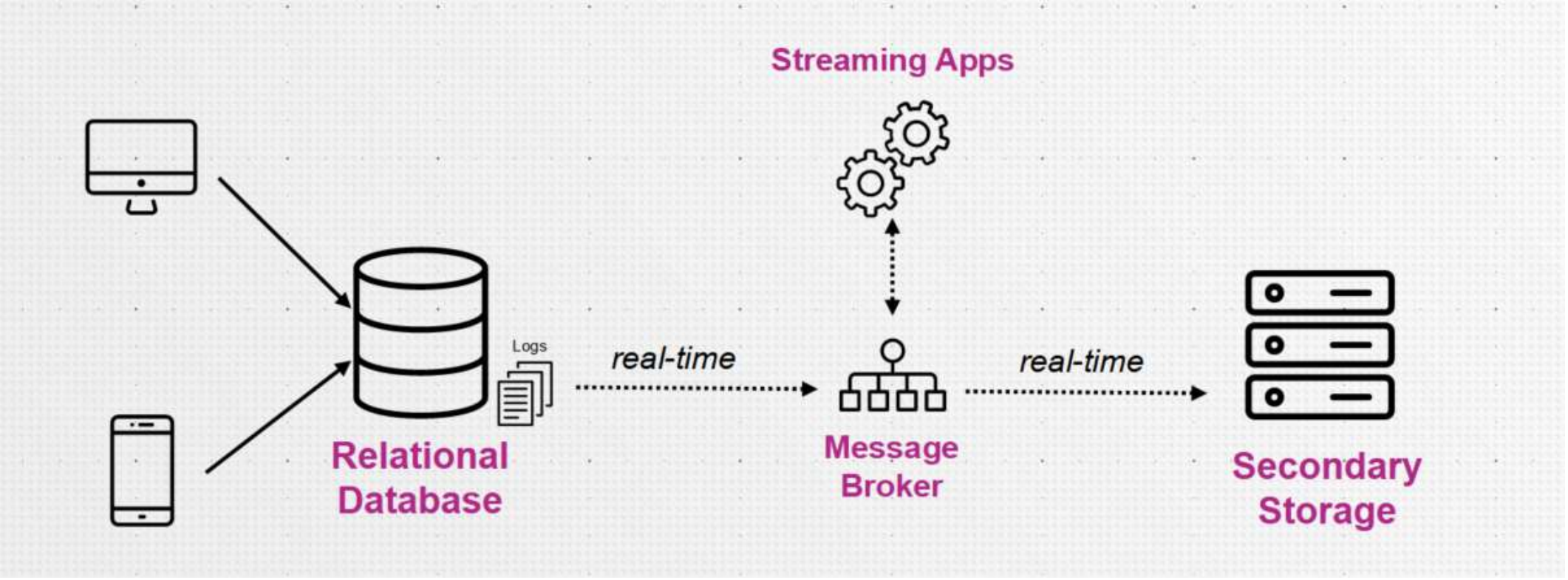
OPCIONES

01	**DATE MODIFIED**	Muchas aplicaciones transaccionales realizan un seguimiento de los metadatos de cada fila, incluyendo quién creó y/o modificó más recientemente la fila, así como cuándo se creó la fila y cuándo se modificó por última vez.
02	**DIFF**	El método diff para la captura de datos de cambios compara el estado actual de los datos con el estado anterior de los datos para identificar qué ha cambiado.
03	**TRIGGERS**	Database triggers pueden utilizarse para realizar CDC en tablas donde se pueda almacenar toda la fila para realizar un seguimiento de cada cambio de columna, o sólo la clave primaria
04	**LOG BASED**	Las bases de datos transaccionales almacenan todos los cambios en un registro de transacciones.

Planteamiento del problema: Replicación en tiempo real



Planteamiento del problema: Replicación en tiempo real - Detalles



Herramientas disponibles

On-prem solutions

ORACLE GoldenGate

IBM InfoSphere software
Change Data Capture

debezium

Qlik Replicate

Cloud-based solutions

AWS DMS
(Data Migrating Services)

GCP Datastream

Azure Data Factory

Casos de Uso (I)

Mainframe Offloading

El mainframe suele utilizarse para el core bancario, pero debido a la capacidad, los costes o las características, puede que necesitemos trasladar los datos a un repositorio diferente.

Application Modernisation

En algunos casos queremos mejorar las aplicaciones existentes con decisiones en tiempo real, analítica, nube, etc. En estos casos CDC es muy conveniente si no quieres tener un impacto en la app existente.

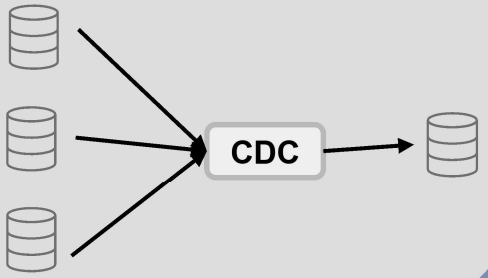
Database Replication

Replicar la base de datos puede ser útil cuando queremos realizar cambios en los datos o en el modelo de datos, con fines analíticos. Es decir, replicar en datawarehaouses, etc.

Casos de Uso (II)

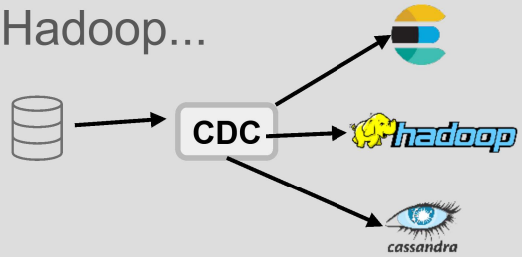
Database Consolidation

Consolidar bases de datos heterogéneas en una única base de datos consolidada.



Multi-store

A partir de un único almacén, distribuya los datos a modernos almacenes de datos independientes como indexadores, NOSQL, Hadoop...



Change Data Capture - Resumen

Change Data Capture (CDC) es una técnica utilizada en bases de datos para identificar y capturar los cambios que ocurren en los datos, como inserciones, actualizaciones o eliminaciones, y luego replicarlos en tiempo real. CDC permite rastrear estos cambios de manera eficiente sin tener que consultar o replicar toda la base de datos, lo que es útil para la integración de datos, análisis en tiempo real, sincronización de bases de datos o la migración de datos.

Esta técnica optimiza el manejo de grandes volúmenes de datos, mejorando la eficiencia en la gestión y análisis.