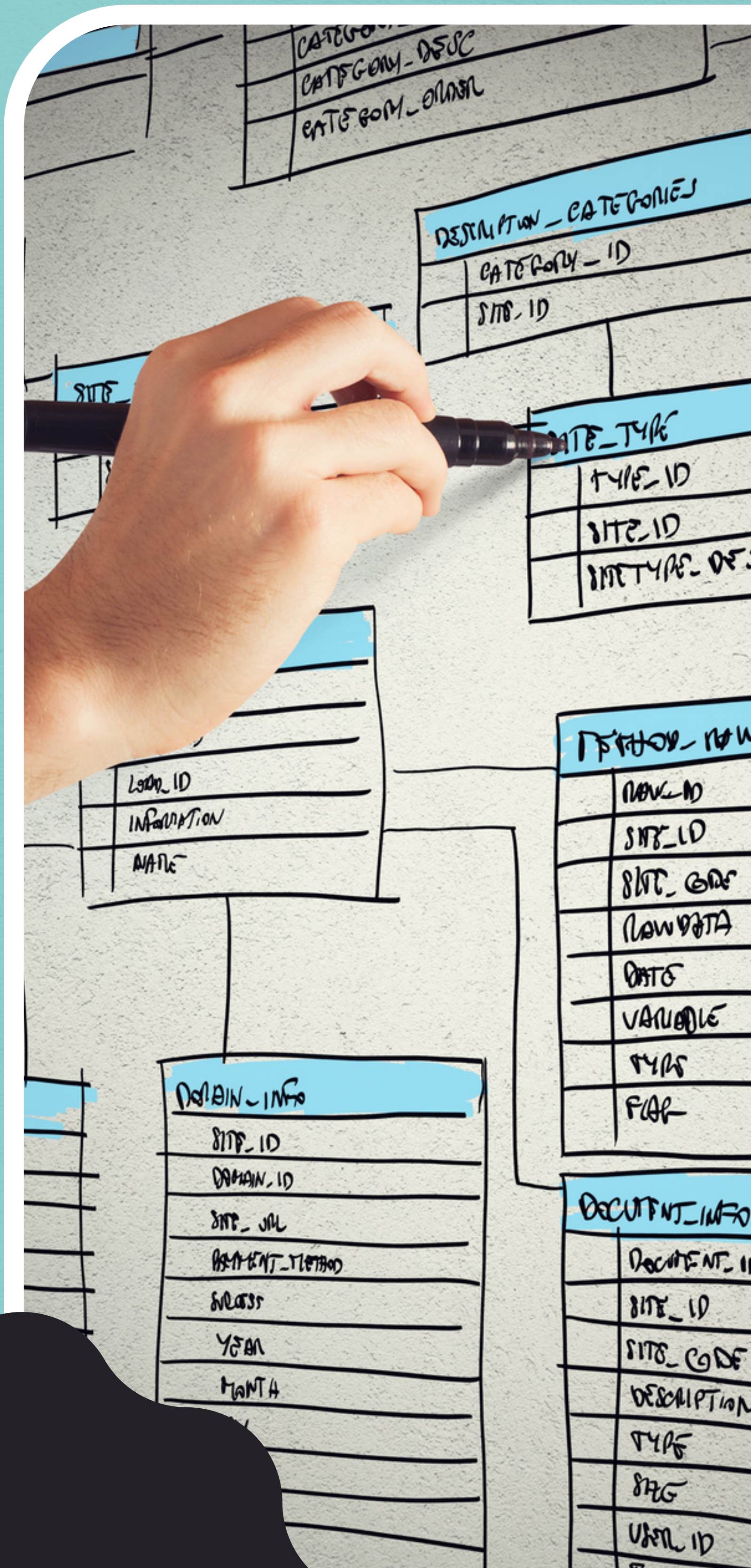
Change Data Capture (CDC) Methods

Various CDC implementation methods have emerged throughout the years. Let's review the most common ones.

in thalia-barrera



The next slides explain the different **CDC methods** used to identify data changes in a source database.

These methods are used in the context of **incremental ETL/ELT**.

●○○○○○○○○

01 table metadata

Using this method requires metadata columns in the source table, such as ***created_at*** or ***updated_at***.

The most common way of ingesting **new and updated rows** in an ETL using this method is to look at the ***updated_at*** column in the destination table to know the latest update and then identify the rows with a later ***updated_at*** in the source table.

Then, the new or updated rows are merged to the destination.

CDC table metadata technique

Source

id	created AT	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-03-2022 20:00
7	01-02-2022 12:00	01-02-2022 12:00
8	01-03-2022 16:30	01-03-2022 16:30

Destination - before replication

id	created AT	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-01-2022 18:00
7	01-02-2022 12:00	01-02-2022 12:00

Max updated_at = 01-02-2022 12:00

Destination - after replication

id	created at	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-03-2022 20:00
7	01-02-2022 12:00	01-02-2022 12:00
8	01-03-2022 16:30	01-03-2022 16:30

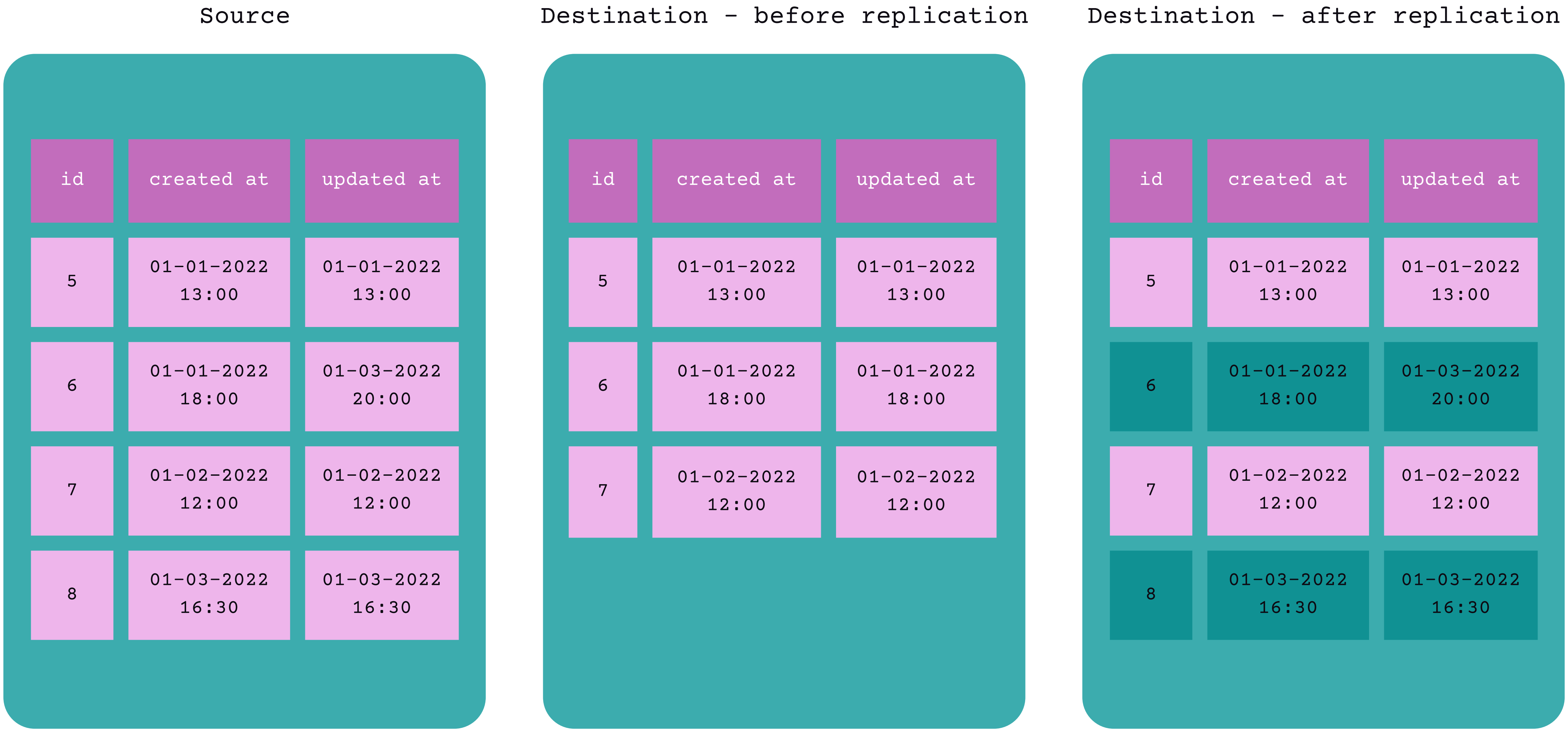
Updated_at (source) > Max updated_at (dest)

02 table differences

This method identifies the difference between the source and the destination tables to detect **new, updated, and even deleted rows**. The difference can be calculated using a **SQL query** or specific utilities provided by the database.

Then, the identified changes are applied to the destination.

CDC table differences technique



Source

id	created at	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-03-2022 20:00
7	01-02-2022 12:00	01-02-2022 12:00
8	01-03-2022 16:30	01-03-2022 16:30

Destination - before replication

id	created at	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-01-2022 18:00
7	01-02-2022 12:00	01-02-2022 12:00

Destination - after replication

id	created at	updated at
5	01-01-2022 13:00	01-01-2022 13:00
6	01-01-2022 18:00	01-03-2022 20:00
7	01-02-2022 12:00	01-02-2022 12:00
8	01-03-2022 16:30	01-03-2022 16:30

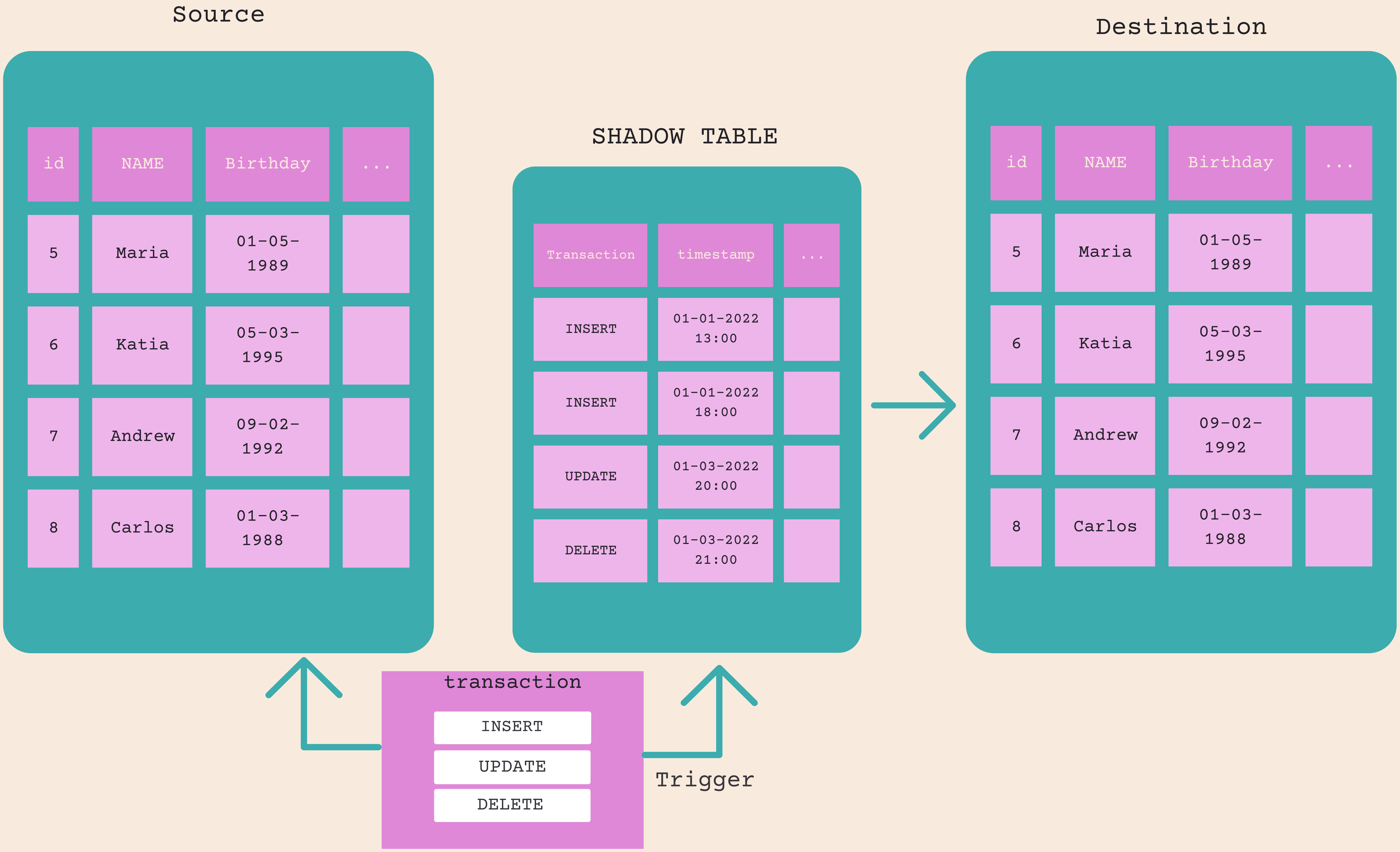
```
SELECT * FROM source EXCEPT
SELECT * FROM destination
```

03 Database triggers (trigger-based CDC)

This method requires the creation of database triggers that execute every time there's an **INSERT, UPDATE or DELETE** operation. The logic in the trigger keeps track of those operations, normally in a separate book-keeping table (often called *shadow table*).

Then, the operations in the shadow table are applied to the destination.

Trigger-based CDC technique



Source

id	NAME	Birthday	...
5	Maria	01-05-1989	
6	Katia	05-03-1995	
7	Andrew	09-02-1992	
8	Carlos	01-03-1988	

SHADOW TABLE

Transaction	timestamp	...
INSERT	01-01-2022 13:00	
INSERT	01-01-2022 18:00	
UPDATE	01-03-2022 20:00	
DELETE	01-03-2022 21:00	

Destination

id	NAME	Birthday	...
5	Maria	01-05-1989	
6	Katia	05-03-1995	
7	Andrew	09-02-1992	
8	Carlos	01-03-1988	

transaction

INSERT

UPDATE

DELETE

Trigger

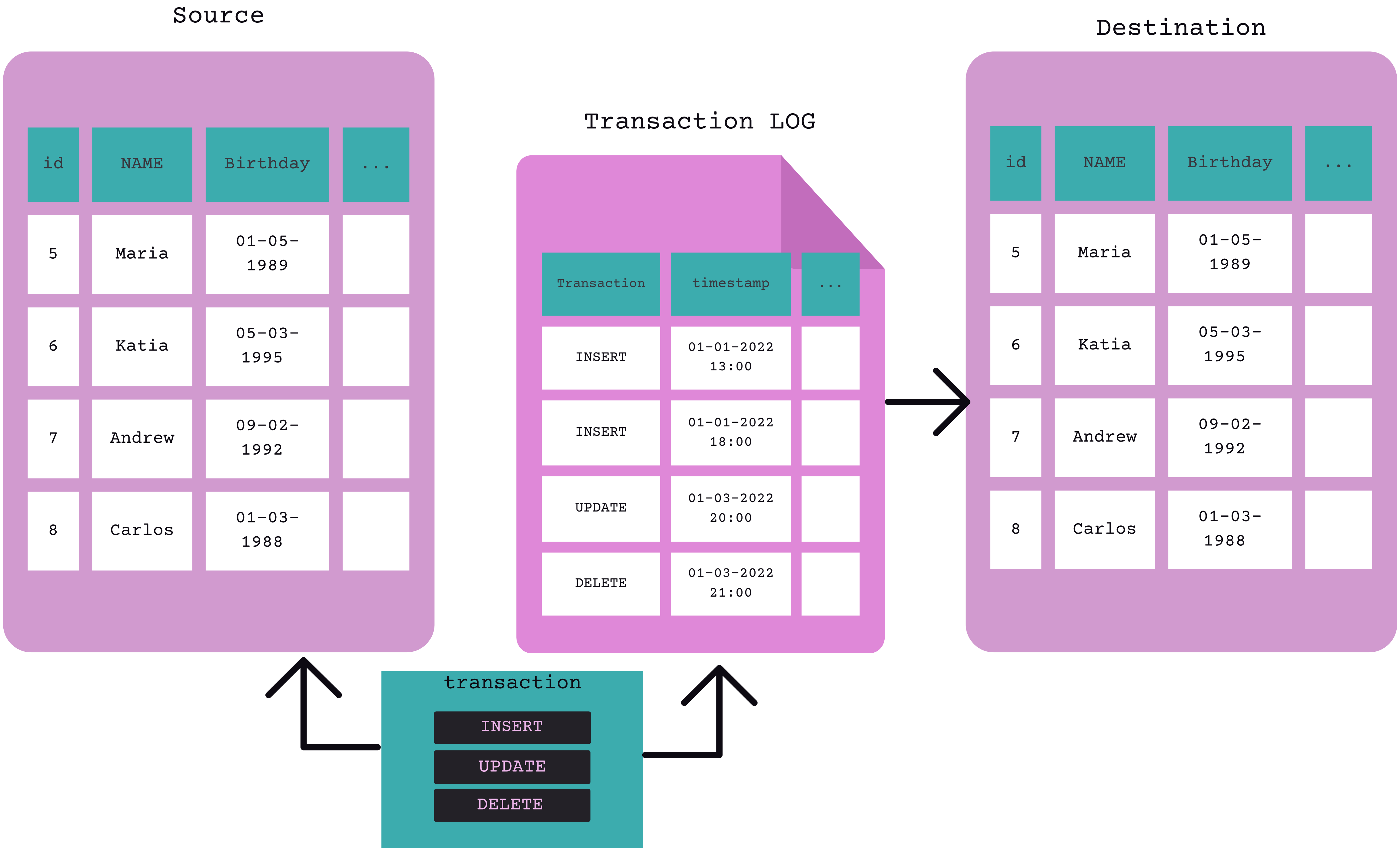
04 Database transaction log (log-based CDC)

Log-based CDC uses the **transaction logs** that some databases - such as Postgres, MySQL, SQL Server, and Oracle - implement natively as part of their core functionality.

Database logs are automatically updated in transactions like **INSERT**, **UPDATE or DELETE**.

Then, the operations in the log are applied to the destination.

Log-based CDC technique



Source

id	NAME	Birthday	...
5	Maria	01-05-1989	
6	Katia	05-03-1995	
7	Andrew	09-02-1992	
8	Carlos	01-03-1988	

Transaction LOG

Transaction	timestamp	...
INSERT	01-01-2022 13:00	
INSERT	01-01-2022 18:00	
UPDATE	01-03-2022 20:00	
DELETE	01-03-2022 21:00	

Destination

id	NAME	Birthday	...
5	Maria	01-05-1989	
6	Katia	05-03-1995	
7	Andrew	09-02-1992	
8	Carlos	01-03-1988	

transaction

INSERT

UPDATE

DELETE

As you can see, there are several approaches for implementing CDC. It's worth mentioning that many modern and **real-time data architectures employ log-based CDC**, which uses a background process to scan database transaction logs for changed data.

Don't forget to
share or save it
for later :)

in thalia-barrera