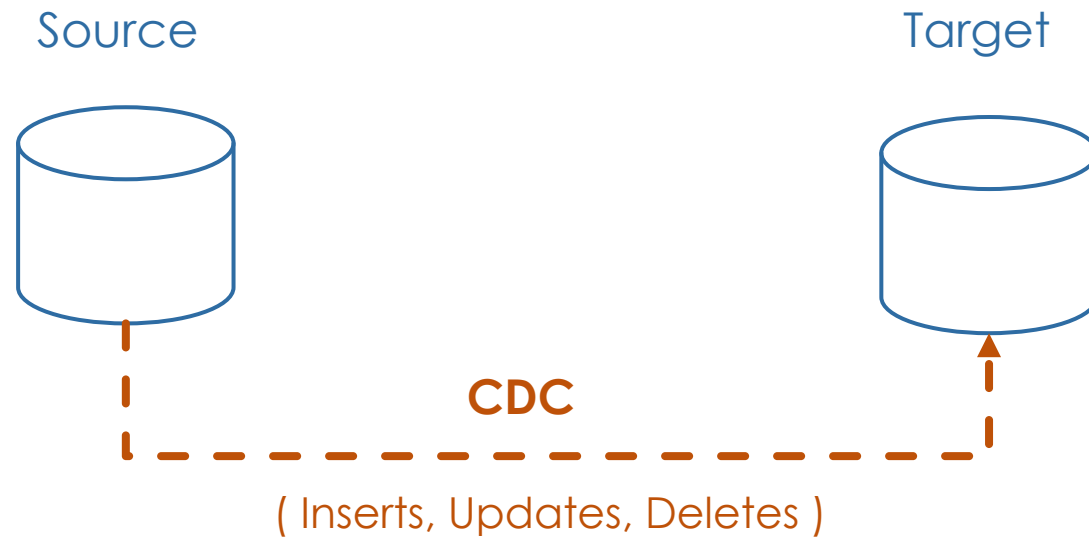# Change Data Capture

Learning Objectives

▶ Definition of Change Data Capture

▶ Processing CDC feed in Delta Lake

Change Data Capture

▶ CDC: Process of identifying changes made to data in the source and delivering those changes to the target

Source

Target

CDC

(Inserts, Updates, Deletes)

Row-Level changes

▶ Inserting new records

▶ Updating existing records

▶ Deleteing existing records

CDC Feed

▶ Row data + metadata

Country ID	Country	Vaccination Rate	sequence	operation
FR	France	0,74	2022-11-01 07:00	UPDATE
FR	France	0.75	2022-11-01 08:00	UPDATE
CA			2022-11-01 00:00	DELETE
US	USA	0,5	2022-11-01 07:00	INSERT
IN	India	0.66	2022-11-01 00:00	INSERT

Country ID	Country	Vaccination Rate
FR	France	0,6
CA	Canada	0,71

CDC Feed

Target table

CDC Feed

▶ Row data + metadata

Country ID	Country	Vaccination Rate	sequence	operation
FR	France	0.74	2022-11-01 07:00	UPDATE
FR	France	0.75	2022-11-01 08:00	UPDATE
CA			2022-11-01 00:00	DELETE
US	USA	0.5	2022-11-01 00:00	INSERT
IN	India	0.66	2022-11-01 00:00	INSERT

Country ID	Country	Vaccination Rate
FR	France	0.75
US	USA	0.5
IN	India	0.66

CDC Feed   Target table

Processing CDC feed

▶ **MERGE INTO** target_table t
   USING source_updates s
   ON t.key_field = s.key_field
   WHEN MATCHED AND t.sequence_field < s.sequence_field
          THEN UPDATE SET *
   WHEN MATCHED AND s.operation_field = "DELETE"
          THEN DELETE
   WHEN NOT MATCHED
          THEN INSERT *

Merge Limitation

▶ Merge can not be performed if multiple source rows matched and attempted to modify the same target row in the Delta table

▶ CDC feed with multiple updates for the same key will generate an exception!

Rank Window Function

▶ `rank().over(Window)`

Country ID	Country	Vaccination Rate	sequence	operation	Rank
FR	France	0.75	2022-11-01 08:00	UPDATE	1
FR	France	0.74	2022-11-01 07:00	UPDATE	2
CA			2022-11-01 00:00	DELETE	1
US	USA	0.5	2022-11-01 00:00	INSERT	1
IN	India	0.66	2022-11-01 00:00	INSERT	1

Rank Window Function

▶ `rank().over(`Window`)`

Partitioning key

Ordering column

Country ID	Country	Vaccination Rate	sequence	operation	Rank
FR	France	0.75	2022-11-01 08:00	UPDATE	1
FR	France	0.74	2022-11-01 07:00	UPDATE	2
CA			2022-11-01 00:00	DELETE	1
US	USA	0.5	2022-11-01 00:00	INSERT	1
IN	India	0.66	2022-11-01 00:00	INSERT	1

Rank Window Function

- `rank().over(Window)`

Country ID	Country	Vaccination Rate	sequence	operation	Rank
FR	France	0.75	2022-11-01 08:00	UPDATE	1
CA			2022-11-01 00:00	DELETE	1
US	USA	0.5	2022-11-01 00:00	INSERT	1
IN	India	0.66	2022-11-01 00:00	INSERT	1