

DISTINCT does not dedupe your table.

D8A.ACADEMY/RESOURCES/DEDUPE-ROW-NUMBER

THE PROBLEM

duplicate	same key, several rows
cause	retries, replays, bad joins
DISTINCT	← only identical rows.

KEEP ONE ROW

ROW_NUMBER()	PARTITION BY the key
ORDER BY	updated_at DESC, keep 1
tie-breaker	← or the winner is random.

FIX THE SOURCE

unique key	the table refuses dupes
MERGE / upsert	update it, do not append
idempotent load	a retry adds nothing

FIND THEM FIRST

GROUP BY key	HAVING COUNT(*) > 1
size the damage	rows vs distinct keys
define the key	what makes it unique

SHORTCUTS

QUALIFY rn = 1	Snowflake, BigQuery
DISTINCT ON (key)	the Postgres way
GROUP BY + MAX	can blend two rows

GOTCHAS

NULL keys	UNIQUE lets them repeat
case and spaces	'a1' is not 'A1'
DELETE in place	stage and check first

WORTH MEMORIZING

01 Keep the latest row per key

Number the rows inside each key, newest first, and keep number one.

```

-- one row per order_id: the latest version
WITH ranked AS (
  SELECT *,
    ROW_NUMBER() OVER (
      PARTITION BY order_id
      ORDER BY updated_at DESC, ingest_id DESC
    ) AS rn
  FROM raw_orders
)
SELECT * FROM ranked WHERE rn = 1;
-- Snowflake, BigQuery: QUALIFY rn = 1, no CTE

```

→ ingest_id is the tie-breaker. Without a unique column at the end of the ORDER BY, two versions with the same updated_at can swap places from one run to the next.

02 Trap: dedupes that do not dedupe

Each one passes a quick look. None of them gives one row per key, every run.

YOU WROTE	WHAT HAPPENS	INSTEAD
DISTINCT *	both versions stay	ROW_NUMBER per key
ORDER BY updated_at	ties flip between runs	add a unique tie-breaker
dedupe in the dashboard	the table still grows	unique key + MERGE

→ Interview phrasing: DISTINCT removes identical rows. Deduplication is choosing one row per business key with a deterministic rule.