Week 2 - Data Transformation & Master Table Creation

Final Master Table

Table Name: `master_table`

Purpose: An integrated dataset that merges key information from six raw datasets including learners, opportunities, cohorts, and campaign data. This table ensures data quality, relational consistency, and readiness for analytics or reporting.

Included Columns:

Column Name	Data Type	Description	Source Table
learner_id	TEXT	Unique identifier of each learner	learner_raw
country	TEXT	Country of the learner	learner_raw
degree	TEXT	Education level	learner_raw
enrollment_id	TEXT	Enrollment reference	learneropportunity_raw
opportunity_id	TEXT	Opportunity unique ID	opportunity_raw
opportunity_name	TEXT	Name of the opportunity	opportunity_raw
category	TEXT	Opportunity category	opportunity_raw
assigned_cohort	TEXT	Cohort ID assigned to learner	learneropportunity_raw
start_date	TIMESTAMP	Cohort start date	cohort_raw
end_date	TIMESTAMP	Cohort end date	cohort_raw
size	INTEGER	Size of the cohort	cohort_raw

Indexes & Constraints:

- `PRIMARY KEY (learner_id)`
- Indexes on `enrollment_id`, `opportunity_id`, `assigned_cohort` for join efficiency
- Text fields normalized using `INITCAP` and `TRIM`

Meaning of the Master Table:

Learner [X] from Pakistan with Degree [Y] enrolled in Opportunity [Z] via Cohort [A]

On Date [D] with cohort size [S]

Table Creation Query

```
DROP TABLE IF EXISTS master_table;

CREATE TABLE master_table (
  learner_id TEXT PRIMARY KEY,
  country TEXT,
  degree TEXT,
  enrollment_id TEXT,
  opportunity_id TEXT,
  opportunity_name TEXT,
  category TEXT,
  assigned_cohort TEXT,
  start_date TIMESTAMP,
  end_date TIMESTAMP,
  size INTEGER
);
```

Stored Procedure Query

```
DROP TABLE IF EXISTS master_table;

CREATE TABLE master_table AS
SELECT
    SPLIT_PART(TRIM(LOWER(lr.learner_id)), '#', 2) AS learner_id,

    INITCAP(TRIM(lr.country)) AS country,
    INITCAP(TRIM(lr.degree)) AS degree,

    SPLIT_PART(TRIM(LOWER(lo.enrollment_id)), '#', 2) AS enrollment_id,

    SPLIT_PART(TRIM(LOWER(orr.opportunity_id)), '#', 2) AS opportunity_id,
    INITCAP(TRIM(orr.opportunity_name)) AS opportunity_name,
    INITCAP(TRIM(orr.category)) AS category,
    lo.assigned_cohort,
```

```
    TO_TIMESTAMP((cr.start_date::DOUBLE PRECISION / 1000)) AS start_date,
    TO_TIMESTAMP((cr.end_date::DOUBLE PRECISION / 1000)) AS end_date,
    cr.size

FROM learner_raw lr

JOIN learneropportunity_raw lo
    ON TRIM(LOWER(lr.learner_id)) = TRIM(LOWER(lo.enrollment_id))

JOIN opportunity_raw orr
    ON TRIM(LOWER(lo.learner_id)) = TRIM(LOWER(orr.opportunity_id))

JOIN cohort_raw cr
    ON TRIM(LOWER(lo.assigned_cohort)) = TRIM(LOWER(cr.cohort_code))

WHERE
    lr.learner_id IS NOT NULL AND
    lo.enrollment_id IS NOT NULL AND
    lo.assigned_cohort IS NOT NULL AND
    orr.opportunity_id IS NOT NULL AND
    cr.cohort_code IS NOT NULL;
```

Data Quality Report

Creating master_table by observing all datasets

1. Issues Detected:

- Redundant rows were **not** found in row-wise duplication checks.
- High **column-level duplicates** in fields like `country`, `degree`, and `institution`.
- Presence of **"Unknown"** or `NULL` values in `degree`, `institution`, and `major`.
- `learner_id`, `enrollment_id`, and `opportunity_id` were often in non-human-readable formats.

Dataset-Level Duplicate Overview

Dataset Name	Total Rows	Duplicate Rows (Exact Matches)
learner_raw	129259	0
learneropportunity_raw	113602	0
cohort_raw	639	0
Cognito_raw	129178	0
Opportunity_raw	187	0
Marketing campaign	143	0

Null Values:

- 71,294 NULLs in opportunity_id, assigned_cohort, apply_date, status, and cohort_size.
- 84,426 NULLs in end_date and start_date.
- 52k+ NULLs in academic attributes like degree, major, and institution.
- 71k records missing opportunity_id, assigned_cohort.
- 84k records missing start and end dates. Significant data gaps in core enrollment attributes.
- All columns (cohort_id, code, start_date, end_date, size) are fully populated.
- Top disciplines include Computer Science (4.7k), Business, Engineering. Major field is moderately complete and well-distributed

Dataset Tables

Dataset	Key Issues Identified	Notes/Actions Required
learner_master	40%+ NULLs in degree, institution, major22k+ NULL in country	Requires imputation and normalization
learner_opportunity	~71k NULLs in assigned_cohort, apply_date, statusDuplicate enrollment_ids found	Clean duplicates and fill missing values
opportunity_raw	tracking_questions column is 100% NULLInconsistent naming	Drop deprecated fields; standardize values
cohort	No NULLsOutliers in cohort_size (e.g., 100,000)	Convert epoch timestamps; cap outliers
cognito	~15–20% NULL in gender, city, stateAge field needs conversion	Derive age from DOB; handle missing values
marketing_campaign	Outliers in amount_spent, reach, results6 duplicate rows	Remove duplicates; Winsorize skewed values

2. Cleaning Logic Applied:

Step	**Method**
Whitespace Handling	`TRIM()`
Capitalization	`INITCAP()`
Filtering Nulls	`WHERE column IS NOT NULL` during inserts
Date Conversion	`TO_TIMESTAMP(column::BIGINT / 1000)`
Column Normalization	Text values cleaned for casing, unknowns excluded

Testing Methodology:

Dataset: master_table

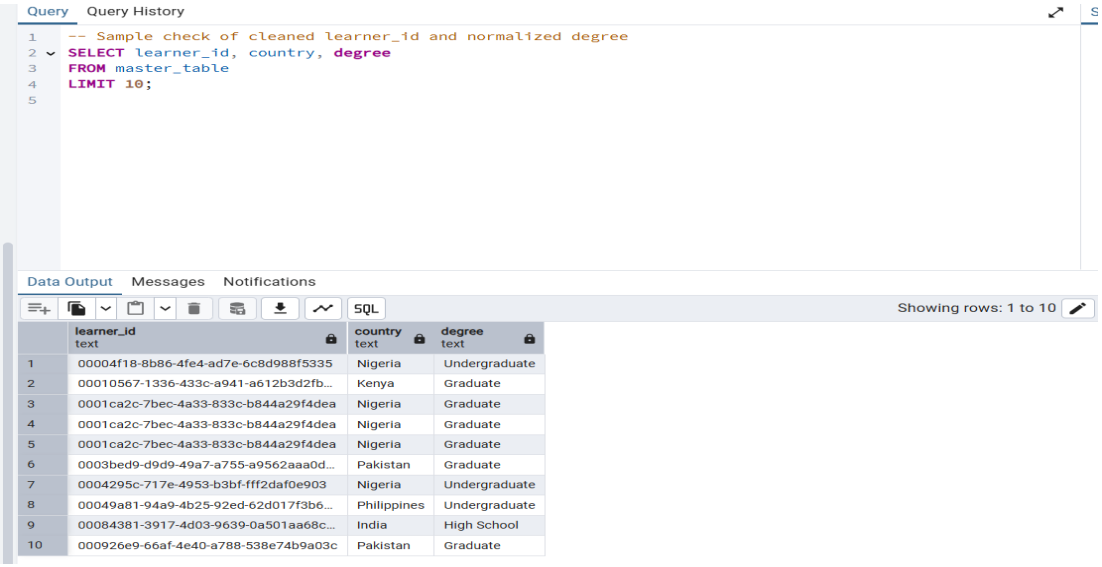
i. Row Count Validation

```
SELECT COUNT(*) AS master_table_row_count FROM master_table;
```

ii. Field-Level Check:

Ensured every field in the final master table had expected value types.
Query:

```
SELECT learner_id, country, degree

FROM master_table

LIMIT 10;
```

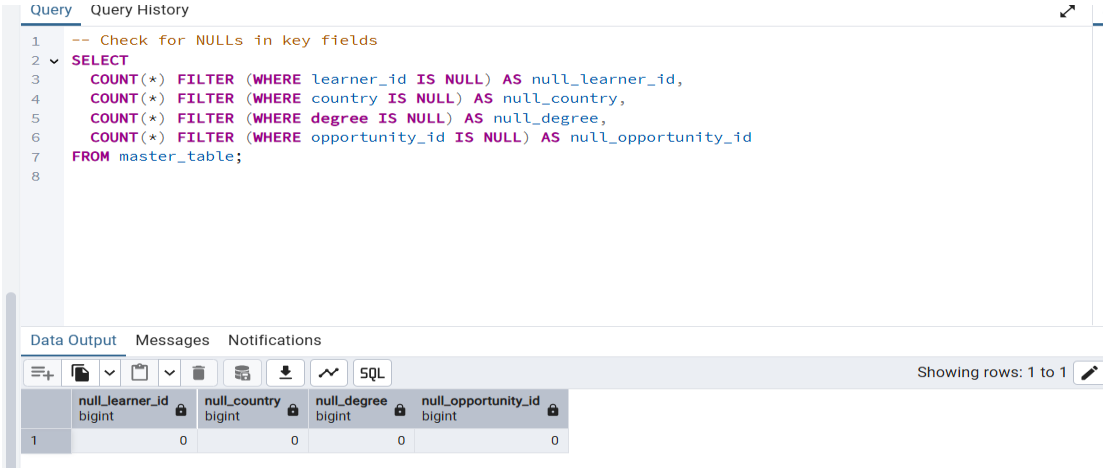


	learner_id text	country text	degree text
1	00004f18-8b86-4fe4-ad7e-6c8d988f5335	Nigeria	Undergraduate
2	00010567-1336-433c-a941-a612b3d2fb...	Kenya	Graduate
3	0001ca2c-7bec-4a33-833c-b844a29f4dea	Nigeria	Graduate
4	0001ca2c-7bec-4a33-833c-b844a29f4dea	Nigeria	Graduate
5	0001ca2c-7bec-4a33-833c-b844a29f4dea	Nigeria	Graduate
6	0003bed9-d9d9-49a7-a755-a9562aaa0d...	Pakistan	Graduate
7	0004295c-717e-4953-b3bf-fff2daf0e903	Nigeria	Undergraduate
8	00049a81-94a9-4b25-92ed-62d017f3b6...	Philippines	Undergraduate
9	00084381-3917-4d03-9639-0a501aa68c...	India	High School
10	000926e9-66af-4e40-a788-538e74b9a03c	Pakistan	Graduate

iii. Verify Null Values:

-- Check for NULLs in key fields

```
SELECT

  COUNT(*) FILTER (WHERE learner_id IS NULL) AS null_learner_id,

  COUNT(*) FILTER (WHERE country IS NULL) AS null_country,

  COUNT(*) FILTER (WHERE degree IS NULL) AS null_degree,

  COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS
null_opportunity_id

FROM master_table;
```

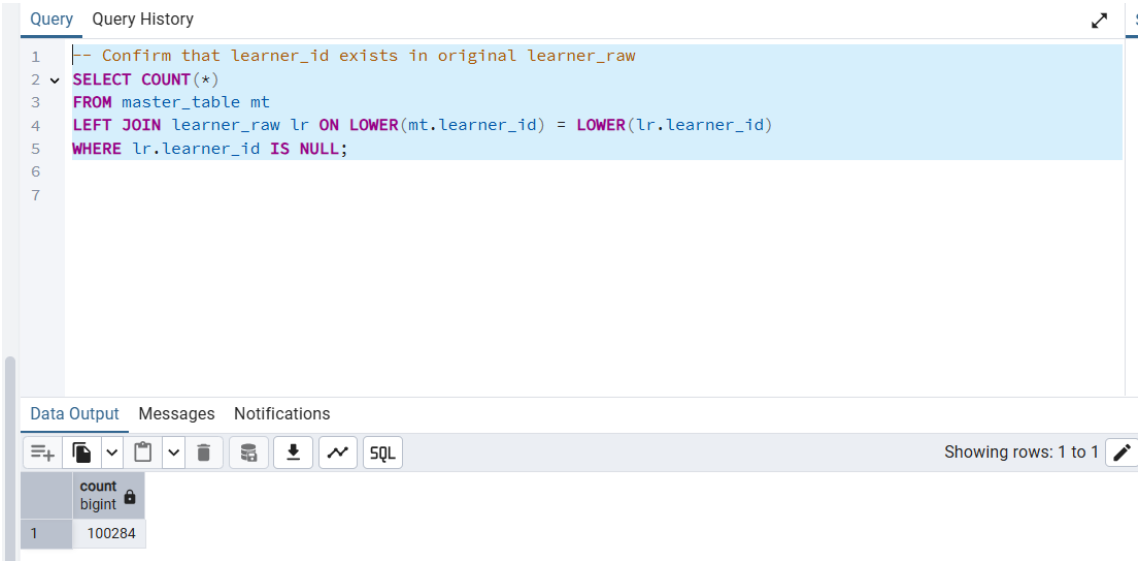


	null_learner_id bigint	null_country bigint	null_degree bigint	null_opportunity_id bigint
1	0	0	0	0

iv. Referential Integrity:

Ensured all join keys matched between datasets.

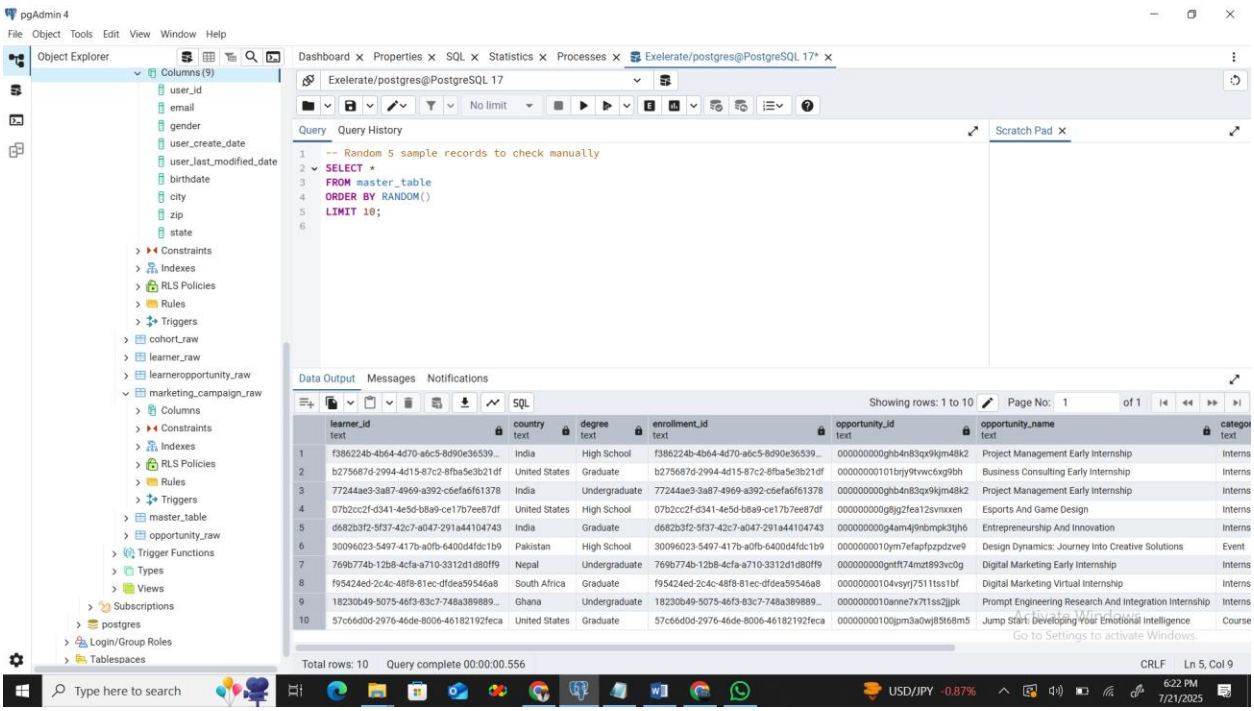
```
-- Confirm that learner_id exists in original learner_raw

SELECT COUNT(*) FROM master_table mt

LEFT JOIN learner_raw lr ON LOWER(mt.learner_id) =
LOWER(lr.learner_id)

WHERE lr.learner_id IS NULL;
```



v. Manual Sampling:

Random records reviewed to verify transformed values.

```
SELECT * FROM master_table ORDER BY RANDOM()

LIMIT 10;
```


pgAdmin 4
File Object Tools Edit View Window Help
Object Explorer
Dashboard Properties SQL Statistics Processes Exelerate/postgres@PostgreSQL 17*
Exelerate/postgres@PostgreSQL 17
Query Query History
Scratch Pad
-- Random 5 sample records to check manually
SELECT *
FROM master_table
ORDER BY RANDOM()
LIMIT 10;
Data Output Messages Notifications
Showing rows: 1 to 10 Page No: 1 of 1
Total rows: 10 Query complete 00:00:00.556
CRLF Ln 5, Col 9
Type here to search
USD/JPY -0.87%
6:22 PM
7/21/2025