

# 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, major 22k+ NULL in country	Requires imputation and normalization
learner_opportunity	~71k NULLs in assigned_cohort, apply_date, status Duplicate enrollment_ids found	Clean duplicates and fill missing values
opportunity_raw	tracking_questions column is 100% NULL Inconsistent naming	Drop deprecated fields; standardize values
cohort	No NULLs Outliers in cohort_size (e.g., 100,000)	Convert epoch timestamps; cap outliers
cognito	~15–20% NULL in gender, city, state Age field needs conversion	Derive age from DOB; handle missing values
marketing_campaign	Outliers in amount_spent, reach, results 6 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;
```

Query Query History

```

1 -- Sample check of cleaned learner_id and normalized degree
2 SELECT learner_id, country, degree
3 FROM master_table
4 LIMIT 10;
5

```

Data Output Messages Notifications

SQL Showing rows: 1 to 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-fff2dafoe903	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;

```

Query Query History

```

1 -- Check for NULLs in key fields
2 SELECT
3 COUNT(*) FILTER (WHERE learner_id IS NULL) AS null_learner_id,
4 COUNT(*) FILTER (WHERE country IS NULL) AS null_country,
5 COUNT(*) FILTER (WHERE degree IS NULL) AS null_degree,
6 COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS null_opportunity_id
7 FROM master_table;
8

```

Data Output Messages Notifications

SQL Showing rows: 1 to 1

	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;
```

The screenshot shows a SQL query editor with a query history tab. The query is as follows:

```
1 -- Confirm that learner_id exists in original learner_raw
2 SELECT COUNT(*)
3 FROM master_table mt
4 LEFT JOIN learner_raw lr ON LOWER(mt.learner_id) = LOWER(lr.learner_id)
5 WHERE lr.learner_id IS NULL;
6
7
```

Below the query editor, the 'Data Output' tab is active, showing a single row of results:

	count bigint
1	100284

The interface also includes a toolbar with icons for file operations, a status bar indicating 'Showing rows: 1 to 1', and tabs for 'Messages' and 'Notifications'.

**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

- Columns (9)
  - user\_id
  - email
  - gender
  - user\_create\_date
  - user\_last\_modified\_date
  - birthdate
  - city
  - zip
  - state
- Constraints
- Indexes
- RLS Policies
- Rules
- Triggers
- cohortLaw
- learner\_raw
- learneropportunity\_raw
- marketing\_campaign\_raw
  - Columns
  - Constraints
  - Indexes
  - RLS Policies
  - Rules
  - Triggers
- master\_table
- opportunity\_raw
- Trigger Functions
- Types
- Views
- Subscriptions
- postgres
- Login/Group Roles
- Tablespaces

Dashboard x Properties x SQL x Statistics x Processes x Exelerate/postgres@PostgreSQL 17\* x

Exelerate/postgres@PostgreSQL 17

Query Query History Scratch Pad

```
1 -- Random 5 sample records to check manually
2 SELECT *
3 FROM master_table
4 ORDER BY RANDOM()
5 LIMIT 10;
```

Data Output Messages Notifications

Showing rows: 1 to 10 Page No: 1 of 1

	learner_id text	country text	degree text	enrollment_id text	opportunity_id text	opportunity_name text	category text
1	f386224b-4b64-4d70-af0c5-bd90e36539...	India	High School	f386224b-4b64-4d70-af0c5-bd90e36539...	000000000ghb4n83qx9km48k2	Project Management Early Internship	Interns
2	b275687d-2994-4d15-87c2-8fba5e3b21df	United States	Graduate	b275687d-2994-4d15-87c2-8fba5e3b21df	000000000101brjy9hwctxg9bh	Business Consulting Early Internship	Interns
3	77244ae3-3a87-4969-a392-cbefa6f61378	India	Undergraduate	77244ae3-3a87-4969-a392-cbefa6f61378	000000000ghb4n83qx9km48k2	Project Management Early Internship	Interns
4	07b2cc2f-d341-4e5d-b8a9-ce17b7ee87df	United States	High School	07b2cc2f-d341-4e5d-b8a9-ce17b7ee87df	000000000g8j2fea12svmaxen	Esports And Game Design	Interns
5	d682b3f2-5f37-42c7-a047-291a44104743	India	Graduate	d682b3f2-5f37-42c7-a047-291a44104743	000000000g4am49nrbmpk3th6	Entrepreneurship And Innovation	Interns
6	30096023-5497-417b-a0fb-6400d4f0c1b9	Pakistan	High School	30096023-5497-417b-a0fb-6400d4f0c1b9	00000000010ym7efapfpzpdzve9	Design Dynamics: Journey Into Creative Solutions	Event
7	769b774b-12b8-4cfa-a710-3312d1d80ff9	Nepal	Undergraduate	769b774b-12b8-4cfa-a710-3312d1d80ff9	000000000gtrf74mzt893cc0g	Digital Marketing Early Internship	Interns
8	f95424ed-2c4c-48f8-81ec-df0ea59546a8	South Africa	Graduate	f95424ed-2c4c-48f8-81ec-df0ea59546a8	000000000104vsyr7511ts1bf	Digital Marketing Virtual Internship	Interns
9	18230b49-5075-46f3-83c7-748a389889...	Ghana	Undergraduate	18230b49-5075-46f3-83c7-748a389889...	00000000010anne7x71ss2gpk	Prompt Engineering Research And Integration Internship	Interns
10	57c66d0d-2976-46de-8006-46182192feca	United States	Graduate	57c66d0d-2976-46de-8006-46182192feca	000000000100pm3adw8586m5	Jump Start: Developing Your Emotional Intelligence	Course

Total rows: 10 Query complete 00:00:00.556

CRLF Ln 5, Col 9

6:22 PM 7/21/2025