MASTER TABLE CREATION

TEAM NAME: 0704 DVA TEAM 34

Team working with the Respective Data Sets

Data Set	Team Member
1. User Data	Sri Hari Ganesh Reddy Konala
2. Learner Opportunity Data	Abeer Asif
3. Cohort Data	Vivek puspalak
4 . Cognito Data	Syed Nazmul Hasan
5. Opportunity Data	V N HARINI
6. Marketing Data	Dinesh Ruthala

Table of Contents:

Introduction..... 4

Phase 1: Understanding & Identifying Data Issues..... 5
Section 1: User Dataset.....5
Section 2: Opportunity Dataset 7
Section 5: Learner Opportunity Dataset.....9
Section 3: Cognito Dataset.....13
Section 4: Cohort Dataset 18
Section 6: Marketing Dataset.....20

Phase 2: Building Master …….....22

Phase 3: Validation & Refinement..... 30

Conclusion.....37

INTRODUCTION

This report presents a comprehensive summary of the work conducted by our team during the data analysis internship project. The primary objective was to construct a robust master table by integrating multiple datasets—Userdata, Cognito, LearnerOpportunity, Opportunity, and Cohort—into a unified, analysis-ready format. This process involved systematic data exploration, preprocessing, and validation steps to ensure high data quality and analytical value.

Throughout the project, our focus remained on identifying and resolving key data issues such as inconsistent formats, missing values, invalid joins, and redundant columns. We performed detailed join operations, standardized null representations, reformatted timestamps, and executed logic-driven row deletions to refine the dataset. The resulting master table serves as a reliable foundation for downstream reporting, modeling, and visualization efforts.

This report outlines the phased approach undertaken, beginning with understanding and documenting individual datasets, followed by constructing and validating the master table. In addition, it highlights the rationale behind each decision, the challenges faced during data integration, and the strategies employed to overcome them. The experience has strengthened our practical skills in data cleaning, transformation, and quality assurance—critical competencies in the field of data analytics.

1. USER DATA:

1. STRUCTURE OF DATA SET:

Column Name	PostgreSQL type
Learner_id	TEXT
Country	TEXT
Degree	TEXT
Institution	TEXT
Major	TEXT

2. PURPOSE OF DATASET:

User data entails information about the level of educational background of learners, v country of study, major undertaken and educational institution.

3. COLUMNS HAVING MISSING DATA:

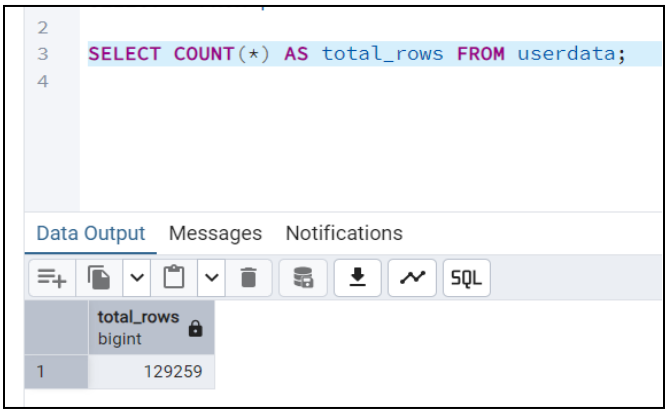
	missing_learner_id bigint	missing_country bigint	missing_degree bigint	missing_institution bigint	missing_major bigint
1	0	2275	52693	52693	52694

4. DUPLICATED RECORDS UNIQUES AND TOTAL ROWS:

- **Unique records:**

	unique_learners bigint	unique_countries bigint	unique_degrees bigint	unique_institutions bigint	unique_majors bigint
1	129259	190	7	34567	4503

• **Total Counts of rows:**



5. INCONSISTENCIES IN FORMATS:

• Make degree, institution, and major follow consistent formatting using title case.

6. . ORPHAN RECORDS:

There are no foreign key hence no orphan record

2. OPPORTUNITY DATA:

Understanding & Identifying Data Issues

1. STRUCTURE OF DATA SET:.

	column_name name	data_type character varying	is_nullable character varying (3)
1	opportunity_id	text	NO
2	opportunity_name	text	NO
3	category	text	NO
4	opportunity_code	text	NO
5	tracking_questions	text	YES

2. PURPOSE OF DATASET:

The purpose of the **Opportunity_Raw(in).csv** dataset is to **store and manage information about various student or participant opportunities**, such as events, competitions, internships, courses, and workshops. These opportunities are likely part of a broader system that tracks engagement, career development, or learning initiatives

3. COLUMNS HAVING MISSING DATA:

Query Query History

```sql
SELECT
    COUNT(*) FILTER (WHERE opportunity_id IS NULL OR opportunity_id = '') AS null_or_empty_id,
    COUNT(*) FILTER (WHERE opportunity_name IS NULL OR opportunity_name = '') AS null_or_empty_name,
    COUNT(*) FILTER (WHERE category IS NULL OR category = '') AS null_or_empty_category,
    COUNT(*) FILTER (WHERE opportunity_code IS NULL OR opportunity_code = '') AS null_or_empty_code,
    COUNT(*) FILTER (WHERE tracking_questions IS NULL OR tracking_questions IN ('', '{}', '""NULL""')) AS null_or_empty_questions,
    COUNT(*) AS total_rows
FROM opportunities;
```

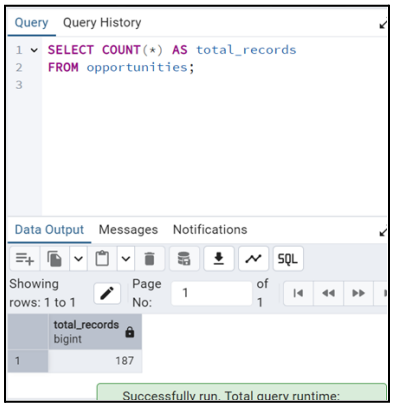
Data Output Messages Notifications

Showing rows: 1 to 1 Page No: 1 of

	null_or_empty_id bigint	null_or_empty_name bigint	null_or_empty_category bigint	null_or_empty_code bigint	null_or_empty_questions bigint	total_rows bigint
1	0	0	0	0	69	187

only tracking_questions having 69 null values.

- **FOR TOTAL RECORDS:**



4. DUPLICATED RECORDS:

No duplicate rows.

5. INCONSISTENCIES IN FORMATS:

Missing tracking_questions: Fill with "Not Provided" or mark explicitly as NULL.

No defined format for names/categories: Apply Title Case standardization.

6. ORPHAN RECORDS:

Will be checked later in learner_oppurtunity data.

3. LEARNER OPPORTUNITY DATA:

Understanding & Identifying Data Issues

1. STRUCTURE OF DATA SET:.

	column_name name	data_type character varying
1	status	integer
2	enrollment_id	text
3	learner_id	text
4	assigned_cohort	character varying
5	apply_date	text

2. PURPOSE OF DATASET:

The learner_opportunity_raw table links learners to specific opportunities and cohorts. It captures when a learner applied (apply_date), their status, and the cohort they were assigned to for a particular opportunity

3. Check for Primary Key Uniqueness:

```
SELECT
   COUNT(*) AS total_records,
   COUNT(DISTINCT enrollment_id) AS unique_enrollment_ids
FROM learner_opportunity_raw;
```

	total_records bigint	unique_enrollment_ids bigint
1	113602	57966

4. Identify Data Quality Issues:

1. Missing values:

	missing_enrollment_id bigint	missing_learner_id bigint	missing_cohort bigint	missing_apply_date bigint	missing_status bigint
1	0	0	13318	188	186

- 13318 missing cohorts
- 188 missing apply dates.
- 186 missing status.

2. Duplicate Records:

	enrollment_id text	count bigint
1	Learner#2444d3b7-3204-4b66-a1e2-72172db26...	3
2	Learner#d7b8c0bd-8fc7-4a9c-a617-369e3fc6ed...	2
3	Learner#6b9871b2-7830-4866-81ad-8674120eb...	2
4	Learner#725951b3-90ed-4494-9bb7-573721cfc...	2
5	Learner#0b248ca5-aa1e-49c6-b447-4b85701d4...	4
6	Learner#b1aa02ae-c3cd-475a-8f3f-a0ebef98113e	2
7	Learner#66cda57a-fac3-4b04-bda9-6ede6a2cb6...	2
8	Learner#a63382a5-b12c-4bdb-9138-f68096a4f2...	4
9	Learner#8ea9690a-b4e4-4a18-a09b-8efe8fd640...	2

There are so many duplicates in enrollment_id

5. INCONSISTENCIES IN FORMATS:

- apply_date is in ISO format string — needs conversion to TIMESTAMP
- status is a float — might need to map to categories or convert to integer
- ID fields (enrollment_id, learner_id) include prefixes that may need stripping

6. ORPHAN RECORDS:

- **Enrollment ID vs. Learner ID**
  The enrollment_id field in the temp_learner_opportunity table logically corresponds to the learner_id in the user_data table. However, due to inconsistent naming conventions and the absence of a foreign key constraint, this relationship is not enforced at the database level.
  Manual validation through SQL joins revealed **186 orphan records.**
  ***Recommendation*:** Standardize column names and implement a foreign key constraint to enforce referential integrity.
- **Opportunity Data Relationship:**The relationship between opportunity_data and learner_opportunity_data is intact.
  ***Observation*: 0 orphan records** were found—indicating a consistent link between opportunities and learner assignments.
- **Cohort Code Consistency:**
  A query checking the consistency between assigned_cohort in temp_learner_opportunity and cohort_code in cohort_data revealed **13,318 orphan records**.
  ***Issue*:** This suggests a significant mismatch, likely due to inconsistent values or missing data in cohort_data

3. Documentation for ETL Planning:

Issue Type	Column	Action Needed
Missing Values	assigned_cohort	Investigate; set as NULL or default if needed
Missing Values	apply_date	Convert or flag NULLs
Missing Values	status	Decide: default, flag, or remove
Format Issue	apply_date	Convert to TIMESTAMP
Format Issue	status	Consider converting float → INT or status mapping
Format Inconsistency	learner_id	Remove "Opportunity#" prefix for joins
Format Inconsistency	enrollment_id	Remove "Learner#" prefix if used for joins

Issue Type	Column	Action Needed
Missing Values	assigned_cohort	Investigate; set as NULL or default if needed
Missing Values	apply_date	Convert or flag NULLs
Duplicates	Entire rows	No duplicates found

4 .COGNITO DATA

Understanding & Identifying Data Issues

1. STRUCTURE OF DATASET

Column Name	Data Type
usercreatedate	timestamp without time zone
userlastmodifieddate	timestamp without time zone
birthdate	date
city	character varying
zip	character varying
user_id	character varying
states	character varying
email	character varying
gender	character varying

2. Preview Sample Data

user_id	email	gender	usercreatedate	userlastmodifieddate	birthdate	city	zip	states
00010567-1336-433c-a941-a312b3d2fbb9	gikonyosalome19@gmail.com	Female	25:58.4	32:50.9	5/4/1998	NAIVASHA	20117	NAKURU
aab0bd67-af03-4e21-816b-101cf05f9a79	evelyn.natasha.guo@gmail.com	NULL	07:08.1	07:20.7	NULL	NULL	NULL	NULL
4656095f-a932-4889-ae96-3b77ff60f1e4	lauren.singh@rocketmail.com	Female	23:54.3	47:51.8	4/5/1990	Queens Village	11428	NY
76b5629f-a024-4de8-9f10-59ebf8fd019b	anihmercy2019@gmail.com	Female	04:21.7	12:28.6	12/28/1998	Ibadan	200221	Oyo
db172006-2017-4b6a-9462-fc2bc7fdfb91	lagnmasame@gmail.com	Female	36:26.4	10:24.5	5/5/1999	Malolos City	3000	Bulacan
78dc6832-dccf-4dab-b7cf-d953acc7c746	koizyinka777@gmail.com	NULL	19:21.4	20:06.2	NULL	NULL	NULL	NULL
2444d3b7-3204-4b68-a1e2-72172db28b33	ujjwal.pandey2103@gmail.com	Male	58:57.6	04:22.0	3/21/2000	New Delhi	110046	Delhi
fec90f6c-de9e-4594-9249-4a5d53ff5a7e	amaliatasabazuing@gmail.com	Female	25:14.7	55:42.2	4/22/2002	Kumasi	233	Ashanti Region
9a4fea59-426a-42e0-ab0d-ed9c0adc6139	230000@d230.org	NULL	33:12.7	32:14.9	NULL	NULL	NULL	NULL
57f7660d-966a-40fd-a1c5-793aed4440f7	vonyedika35@gmail.com	NULL	58:37.4	57:57.2	NULL	NULL	NULL	NULL

3. Check for Primary Key Uniqueness

Query Query History

```sql
SELECT COUNT(*) AS total_rows, COUNT(DISTINCT user_id) AS unique_user_ids
FROM cognitodata;
```

Data Output Messages Notifications

	total_rows bigint	unique_user_ids bigint
1	129178	129178

Findings:

user_id is the strongest candidate for a Primary Key (PK) — every entry is unique and complete.

4. Identify Data Quality Issues:

- **Missing Values**

Query Query History

```sql
SELECT
  COUNT(*) AS total_rows,
  COUNT(email) AS email_not_null,
  COUNT(*) - COUNT(email) AS email_missing,
  COUNT(gender) AS gender_not_null,
  COUNT(*) - COUNT(gender) AS gender_missing,
  COUNT(city) AS city_not_null,
  COUNT(*) - COUNT(city) AS city_missing,
  COUNT(zip) AS zip_not_null,
  COUNT(*) - COUNT(zip) AS zip_missing,
  COUNT(states) AS states_not_null,
  COUNT(*) - COUNT(states) AS states_missing
FROM cognitodata;
```

Data Output Messages Notifications

	total_rows bigint	email_not_null bigint	email_missing bigint	gender_not_null bigint	gender_missing bigint	city_not_null bigint	city_missing bigint	zip_not_null bigint	zip_missing bigint	states_not_null bigint	states_missing bigint
1	129178	129178	0	86316	42862	86315	42863	86311	42867	86314	42864

Findings:

- There are a number of missing values in gender, city, zip and states columns

- **Duplicate Email Records**

Query Query History

```
SELECT email, COUNT(*) AS occurre
FROM cognitodata
GROUP BY email
HAVING COUNT(*) > 1;
```

Data Output Messages Notifications

	email character varying (255)	occurrences bigint
1	daliafouad12345@gmail.com	2
2	nehalmallik00@gmail.com	2
3	minahilmumtaz54@gmail.com	2
4	knnabuihe@gmail.com	2
5	amaji5295@gmail.com	2
6	wediaishaq17@gmail.com	2
7	amrita02051999@gmail.com	2
8	mwirigikennedy820@gmail.com	2
9	bibilola77@gmail.com	2

Findings:

- There are 8 emails that have been used more than once

5. Inconsistent City/State Formatting:

Inconsistent Cities	Inconsistent State
Abakailiki	abia
Abakakili	Abia
Abakaliki	ABIA
Accar	Abia s
Acccra	abia state
Accea	Abia state
Acceq	Abia State
accra	ABIA STATE
Accra	ablekuma
ACCRA	Ablekuma
addis ababa	Abu dabhi
Addis Ababa	abu dhabi
ADDIS ABABA	Abu dhabi
addis abeba	Abu Dhabi
Addis abeba	Abu Dhabi uae

Findings:

1. **Inconsistent Naming of States and Cities:**
   - Multiple variations in the spelling of **Abu Dhabi** (e.g., *Abu dabhi, abu dhabi, Abu Dhabi uae*).

    - Lack of standardization in capitalization (e.g., *abu dhabi* vs. *Abu Dhabi*).
    - Inclusion of extra descriptors such as “UAE” in some entries.
2. **Misspellings and Typos:**
    - Various locations with spelling errors or incorrect characters.
    - Inconsistent representations due to minor spelling mistakes.
3. **Extra White Spaces:**
    - Leading or trailing spaces causing discrepancies.
    - Multiple spaces between words affecting data consistency.

- **Future Birthdates:**

Findings:

- There are no birthdays that haven’t happened yet

- **Creation Date After Modification Date**

6. Document Issues & Plan for ETL

Issue Type	Column(s)	Observation	Suggested ETL Fix
Missing Data	gender, city, zip, states	Concerning number of missing values	Default to "unknown" or flag for review
Duplicate Records	email	Duplicate emails found	Retain latest by UserLastModifiedDate
Format Inconsistency	city, states	Inconsistent Naming, Misspellings, Typos or extra white spaces	Required normalization
Format Issues	zip	Non-standard formats	Normalize or flag

5. COHORT DATA:

Understanding & Identifying Data Issues:

1. STRUCTURE OF DATASET:

Column Name	PostgreSQL type
Cohort_id	TEXT
Cohort_code	TEXT
Start_date	bigint
End_date	bigint
size	bigint

2. PURPOSE OF DATASET:

The cohort_data dataset is designed to monitor groups of learners enrolled in particular educational programs over a defined time period. Each cohort is characterized by:

- A unique identifier (cohort_id, cohort_code)
- The number of learners in the group (cohort size)
- Start and end dates indicating the duration of the cohort.

TOTAL RECORDS:

- Total Records: 639 rows

3. Identify Data Quality Issues:

```
-- Check for NULLs
SELECT COUNT(*) FROM cohortraw WHERE cohort_id IS NULL OR cohort_code IS NULL OR size IS NULL;

-- Check for duplicates
SELECT cohort_id, cohort_code, COUNT(*) FROM cohortraw
GROUP BY cohort_id, cohort_code
HAVING COUNT(*) > 1;
```

- **MISSING VALUES:**

- Missing Values: None. All columns are fully populated.
- Duplicate Records: 0 duplicate rows detected.

4. INCONSISTENCIES IN FORMATS:
The cohort_code contains a Short alphanumeric identifier so keep its datatype VARCHAR(20) instead of text.

• As the cohort_id has no content, check if the cohort_code is unique so we could set it as a primary key and drop cohort_id column

. • As start_date and end_date is in text, convert it into DATE datatype or TImeStamp.

5. ORPHAN RECORDS: Right now there is no link between Cohort and Learners.

6. MARKETING DATA:

Understanding & Identifying Data Issues:

1. **STRUCTURE OF DATASET :**

COLUMN NAME	DETECTED DATATYPE
Ad Account Name	TEXT
Campaign name	TEXT
Delivery status	TEXT
Delivery level	TEXT
Reach	INTEGER
Outbound clicks	INTEGER
Landing page views	INTEGER
Result type	TEXT
Results	INTEGER
Cost per result	NUMERIC (FLOAT)
Amount spent (AED)	NUMERIC (FLOAT)
CPC (cost per link click)	NUMERIC (FLOAT)
Reporting starts	DATE
rpc	NUMERIC (FLOAT)

2. **PURPOSE OF DATASET :**

The marketing campaign dataset is designed to analyze the performance and efficiency of digital advertising efforts. It helps track key metrics such as reach, clicks, conversions, and cost-related figures. This enables marketers to evaluate campaign effectiveness, optimize future strategies, and make data-driven decisions for better return on investment (ROI)

3. Identify Data Quality Issues:

1. Missing Values

- Check for NaN or blanks:
- `rpc` has missing values in some records.

2. Duplicate Records

Duplicates may occur if the same campaign is reported multiple times.

4. Inconsistent Formats:

- Campaign names use inconsistent prefixing like #, ##, or ###.
- Values like `Reach`, `Cost`, `Results` are floats/integers—may need standard rounding.
- Dates are uniform (`YYYY-MM-DD`) – good.

5. Document Findings for ETL Planning

Issue	Column(s)	Transformation Required
Missing values	`rpc`	Fill with derived values or flag for review
Inconsistent campaign naming	`Campaign name`	Standardize campaign names (remove #, use proper casing)
Outliers in performance metrics	`Reach`, `Results`, `rpc`	Flag records that are statistically extreme
Duplicate campaign entries	`Campaign name`	Remove or deduplicate based on highest `Results` or latest
Mixed format in naming prefixes	`Campaign name`	Remove or normalize prefixing (#, ##, ###)

Overview:

This document summarizes the full data cleaning lifecycle applied to the master_table, which was created by joining and integrating multiple sources: **userdata, cognito, learneropportunity, Opportunity, and master_cohort.** The goal was to produce a reliable, analysis-ready table for downstream reporting, modeling, and visualization.

Initial Problems Identified:

1. Null Values in Critical Columns:

opportunity_id, assigned_cohort, opportunity_name, opportunity_code, and several cohort fields contained high NULL counts.

2. Text-Based Nulls:

Many columns contained 'null', 'NULL', 'none', and blank strings as text, which PostgreSQL doesn't recognize as real NULLs.

3. Date Format Issues:

Epoch milliseconds needed conversion into timestamp fields for columns like start_date, end_date, and apply_date.

4. Missing Key Fields:

Joins failed for rows with missing learner_id, opportunity_id, or cohort_code.

5. Unused Columns:

tracking_questions had high NULLs and was deemed unnecessary for final analysis.

Master Table Join Summary:

Join With Table	Why We Join It	What Info It Adds
Userdata (Base)	Main learner details	Country, degree, institution, major
Cognito	Identity data (via user_id)	Email, gender, location, birthdate
LearnerOpportunity	Applications history	Cohort applied to, application date, status
Opportunity	Opportunity metadata	Category, name, code, tracking info
Cohort	Cohort structure	Start/end date, duration, size
MarketingCampaign (optional)	Campaign performance	Clicks, reach, cost per application (join not direct — aggregate separately)

Primary Key (PK) and Foreign Key (FK) Summary:

Table	Primary Key (PK)	Foreign Key (FK)
Userdata	learner_id	—
Cognito	user_id	— (joined to Userdata.learner_id after strip)
LearnerOpportunity	enrollment_id	learner_id, assigned_cohort
Opportunity	opportunity_id	Can join with LearnerOpportunity if opportunity assigned is available
Cohort	cohort_code	—

- **Found Issue of inappropriate column names, enrollment_id (should be learner_id)**

Before:

Query Query History

```
SELECT * FROM public.learneropportunity
```

Data Output Messages Notifications

Showing r

	enrollment_id text	learner_id text	assigned_cohort text	apply_date text	status integer
1	Learner#4e6f78a9-f9b2-4352-ad22-d43dc46f5ff7	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-10T06:28:31.902Z	1070
2	Learner#4e79d245-3436-4fec-9906-901a03639a...	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2023-11-15T03:08:17.442Z	1120
3	Learner#4e9f5cb5-0576-4dbc-b7f5-1fae5f29b2df	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-06T14:07:01.322Z	1070
4	Learner#4ea61aa9-17da-4b60-9872-359b8e1e16...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-04-11T22:01:13.548Z	1070
5	Learner#4eb218c7-467a-470a-9e3e-a2b7bc649e...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-10-22T15:44:13.402Z	1120
6	Learner#4ec728db-7d09-4a8e-b1ab-dab3011a55...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-04-12T07:00:26.574Z	1070
7	Learner#4edc150c-ea73-4144-993f-77ca8124de...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-11-24T11:07:11.742Z	1070
8	Learner#4ef3715a-e420-4296-908c-eab9e5b797...	Opportunity#0000000010WCBS50CYGDX97ES4	BC69M2K	2025-02-18T12:41:51.125Z	1070
9	Learner#4ef49684-e9b0-40a9-b07e-07bfa78bdbc5	Opportunity#0000000010WCBS50CYGDX97ES4	BGRQZ2N	2024-10-08T09:50:06.608Z	1120
10	Learner#4f027693-d86f-4d65-a0a1-6342c8c0979e	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-04T15:15:45.184Z	1070

Resolving Identified Issues:

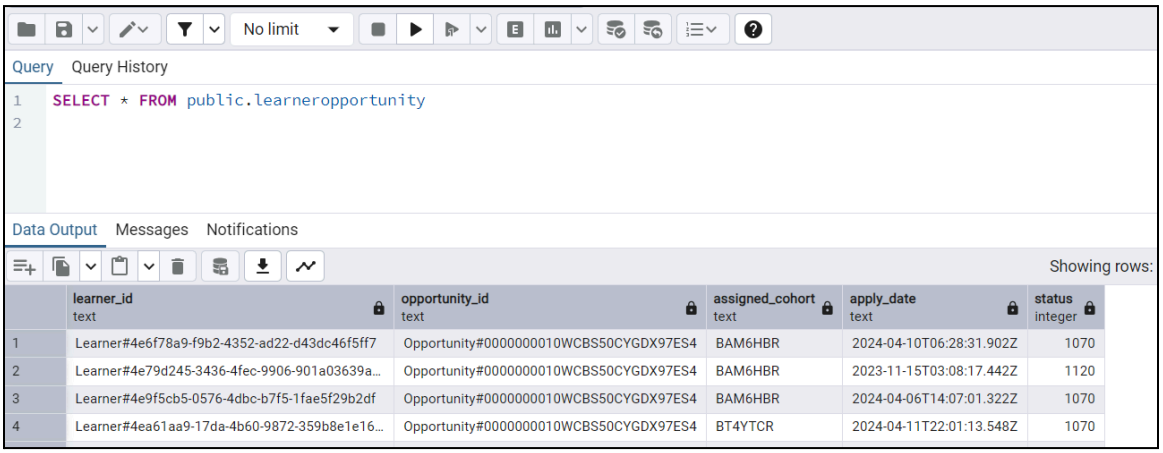
Learner_id to be opportunity_id

```
-- Step 1: Temporarily rename learner_id to something else
ALTER TABLE public.learneropportunity
RENAME COLUMN learner_id TO temp_opportunity_id;

-- Step 2: Rename enrollment_id to learner_id
ALTER TABLE public.learneropportunity
RENAME COLUMN enrollment_id TO learner_id;

-- Step 3: Rename temp_opportunity_id to opportunity_id
ALTER TABLE public.learneropportunity
RENAME COLUMN temp_opportunity_id TO opportunity_id;
```

Data Output Messages Notifications

```
ALTER TABLE

Query returned successfully in 41 msec.
```

After resolving the issue:



Query Query History

```
SELECT * FROM public.learneropportunity
```

Data Output Messages Notifications

Showing rows:

	learner_id text	opportunity_id text	assigned_cohort text	apply_date text	status integer
1	Learner#4e6f78a9-f9b2-4352-ad22-d43dc46f5ff7	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-10T06:28:31.902Z	1070
2	Learner#4e79d245-3436-4fec-9906-901a03639a...	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2023-11-15T03:08:17.442Z	1120
3	Learner#4e9f5cb5-0576-4dbc-b7f5-1fae5f29b2df	Opportunity#0000000010WCBS50CYGDX97ES4	BAM6HBR	2024-04-06T14:07:01.322Z	1070
4	Learner#4ea61aa9-17da-4b60-9872-359b8e1e16...	Opportunity#0000000010WCBS50CYGDX97ES4	BT4YTCR	2024-04-11T22:01:13.548Z	1070

Key Relationships (Joins)

From Table	Join Column	To Table	Join Column
userdata	learner_id	cognito	REPLACE(user_id)
userdata	learner_id	learneropportunity	learner_id
learneropportunity	opportunity_id	"Opportunity"	opportunity_id
learneropportunity	assigned_cohort	master_cohort	cohort_code

<u>MASTER TABLE:</u>

ER Table for Master Table Structure:

Table Name	Column Name	Key Type	Description
userdata	learner_id	PK	Unique identifier for each learner
userdata	country		Country of the learner
userdata	degree		Degree type (e.g., undergraduate)
userdata	institution		Institution name
userdata	major		Major or specialization
cognito	user_id	PK	Matches `learner_id` after stripping 'Learner#'
cognito	email		Email address
cognito	gender		Gender of learner

cognito	usercreateddate		Account creation time
cognito	userlastmodifieddate		Last update time
cognito	birthdate		Date of birth
cognito	city		City of residence
cognito	zip		Zip code
cognito	state		State/Province
learneropportunity	learner_id	FK → userdata	Links to learner
learneropportunity	opportunity_id	FK → Opportunity	Links to opportunity
learneropportunity	assigned_cohort	FK → master_cohort	Cohort the learner is assigned to
learneropportunity	apply_date		When the learner applied
learneropportunity	status		Application status
Opportunity	opportunity_id	PK	Unique ID of the opportunity
Opportunity	opportunity_name		Title of the opportunity
Opportunity	category		Type (e.g., internship, event)
Opportunity	opportunity_code		Code identifier
Opportunity	tracking_questions		Optional tracking metadata
master_cohort	cohort_code	PK	Code identifying the cohort
master_cohort	start_date		Cohort start timestamp
master_cohort	end_date		Cohort end timestamp
master_cohort	size		Number of learners in the cohort
master_cohort	duration_days		Derived from (end_date - start_date)

Master Table Query:

```
DROP TABLE IF EXISTS public.master_table;

CREATE TABLE public.master_table AS

SELECT

    -- USERDATA

    u.learner_id,

    u.country,

    u.degree,

    u.institution,

    u.major,


    -- COGNITO (account data)

    c.email,

    c.gender,

    c.usercreatedate,

    c.userlastmodifieddate,

    c.birthdate,

    c.city,

    c.zip,

    c.state,


    -- LEARNER OPPORTUNITY

    lo.opportunity_id,

    lo.assigned_cohort,

    lo.apply_date,
```

```
    lo.status,
    -- OPPORTUNITY DETAILS
    o.opportunity_name,
    o.category,
    o.opportunity_code,
    o.tracking_questions,

    -- COHORT DETAILS
    mc.cohort_code,
    mc.start_date,
    mc.end_date,
    mc.size AS cohort_size,

    -- DURATION in days
    ROUND(EXTRACT(EPOCH FROM (mc.end_date - mc.start_date)) / 86400)
AS duration_days

FROM public.userdata u

-- Match learner_id to learneropportunity
LEFT JOIN public.learneropportunity lo
    ON u.learner_id = lo.learner_id

-- Match with Opportunity table
LEFT JOIN public."Opportunity" o
```

```
    ON lo.opportunity_id = o.opportunity_id

-- Join with Cognito (strip 'Learner#' from learner_id)
LEFT JOIN public.cognito c
    ON c.user_id = REPLACE(u.learner_id, 'Learner#', '')

-- Join with master_cohort
LEFT JOIN public.master_cohort mc
    ON lo.assigned_cohort = mc.cohort_code;
```

DATA CLEANING PROCESS:

- **Total Null Values Count from the Master Table**

```sql
SELECT * FROM public.master_table

-- Total 'null', 'NULL', 'None', '', etc. before cleaning for key fields
SELECT
  COUNT(*) FILTER (WHERE TRIM(LOWER(tracking_questions)) IN ('null', 'none', '')) AS tracking_questions_nulls,
  COUNT(*) FILTER (WHERE TRIM(LOWER(apply_date)) IN ('null', 'none', '')) AS apply_date_nulls,
  COUNT(*) FILTER (WHERE TRIM(LOWER(institution)) IN ('null', 'none', '')) AS institution_nulls,
  COUNT(*) FILTER (WHERE TRIM(LOWER(major)) IN ('null', 'none', '')) AS major_nulls,
  COUNT(*) FILTER (WHERE TRIM(LOWER(degree)) IN ('null', 'none', '')) AS degree_nulls,
  COUNT(*) FILTER (WHERE TRIM(LOWER(gender)) IN ('null', 'none', '')) AS gender_nulls,
  COUNT(*) FILTER (WHERE assigned_cohort IS NULL) AS assigned_cohort_nulls,
  COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS opportunity_id_nulls,
  COUNT(*) FILTER (WHERE duration_days IS NULL) AS duration_days_nulls
FROM public.master_table;
```

Data Output Messages Notifications

Showing rows: 1 to 1 Page

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint
1	0	0	52966	52929	52707	0	84426	71294	84426

- **Resolving the issue: Standardize Nulls (Text to Real NULL)**

```sql
UPDATE public.master_table
SET
  tracking_questions = NULLIF(TRIM(tracking_questions), ''),
  category = NULLIF(TRIM(category), ''),
  opportunity_name = NULLIF(TRIM(opportunity_name), ''),
  opportunity_code = NULLIF(TRIM(opportunity_code), ''),
  apply_date = NULLIF(TRIM(apply_date), ''),
  institution = NULLIF(TRIM(institution), ''),
  degree = NULLIF(TRIM(degree), ''),
  major = NULLIF(TRIM(major), ''),
  city = NULLIF(TRIM(city), ''),
  state = NULLIF(TRIM(state), ''),
  zip = NULLIF(TRIM(zip), ''),
  gender = NULLIF(TRIM(gender), ''),
  email = NULLIF(TRIM(email), '')
```

Data Output Messages Notifications

```
UPDATE 53344

Query returned successfully in 9 secs 583 msec.
```

- **Counting the Actual Nulls (After Cleaning)**

```sql
-- Count actual NULLs after cleaning
SELECT
  COUNT(*) FILTER (WHERE tracking_questions IS NULL) AS tracking_questions_nulls,
  COUNT(*) FILTER (WHERE apply_date IS NULL) AS apply_date_nulls,
  COUNT(*) FILTER (WHERE institution IS NULL) AS institution_nulls,
  COUNT(*) FILTER (WHERE major IS NULL) AS major_nulls,
  COUNT(*) FILTER (WHERE degree IS NULL) AS degree_nulls,
  COUNT(*) FILTER (WHERE gender IS NULL) AS gender_nulls,
  COUNT(*) FILTER (WHERE assigned_cohort IS NULL) AS assigned_cohort_nulls,
  COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS opportunity_id_nulls,
  COUNT(*) FILTER (WHERE duration_days IS NULL) AS duration_days_nulls
FROM public.master_table;
```

Data Output Messages Notifications

Showing rows: 1 to 1 Page No:

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint
1	86471	71296	0	0	0	43013	84426	71294	84426

- **Updating the NULL values in gender column as “unknown”**

```sql
UPDATE public.master_table
SET gender = 'Unknown'
WHERE gender IS NULL;
```

Data Output Messages Notifications

```
UPDATE 43013

Query returned successfully in 1 secs 619 msec.
```

- **The master_table is Still Having Nulls, so we are cleaning again.**

```
SELECT * FROM public.master_table
```

Data Output Messages Notifications

Showing rows: 1 to 1000 Page No: 1 of 185

	learner_id text	country text	degree text	institution text	major text
21	Learner#6473392e-b871-4e37-9b51-d9a17a40feeb	Nigeria	Graduate Student	Esut	Accounting
22	Learner#64dc7392-fb79-482c-8e91-fc5266c31a7c	Nigeria	High School Student	Government scho	Engineering
23	Learner#66816a9d-db4e-4c03-a34b-5a29f616fe15	Philippines	Graduate Student	Don Mariano Marcos Memorial State University	Hospitality Management
24	Learner#67d38120-0433-4640-adfe-86801c023f7a	Philippines	Not in Education	N/A	Othe
25	Learner#b09561fc-3d4c-4a95-9404-933233f46a7a	Nigeria	Not in Education	Imo State University Owerri	Psychology
26	Learner#b0c2ae19-7b3a-44d4-9848-38335e6c925f	India	High School Student	HSC	Other
27	Learner#b211d25c-8685-43a4-b34b-61e6153e45...	Egypt	Graduate Student	Marwan Mohsen	Social Studies
28	Learner#b3f414c8-858d-467e-b99a-56d1ec2049e7	Kenya	Undergraduate Student	Jomo Kenyatta University of Agriculture and Technology	Project Management
29	Learner#b416a2f1-5ce5-40e2-be1b-2f73fe6f599b	Nigeria	High School Student	Caleb international college	Computer Science
30	Learner#b5505414-47b1-47a3-9f4c-d163f6202e6d	India	Undergraduate Student	Tvnmc	Medicine
31	Learner#7cd12df8-1f13-4fb2-83e8-e76452efceeb	Guinea	NULL	NULL	NULL
32	Learner#b70eec08-622e-465e-aa30-613a5804ffbe	United States	Undergraduate Student	Oak Hills High School	Medical Science
33	Learner#b7d334c2-387b-461c-8fd5-7b8a9531f22e	Nigeria	Graduate Student	Stratford University	Computer Information System
34	Learner#b85043be-ffcd-439d-bbf7-161f970a609e	Pakistan	Other Professional	Saira Miraj memorial hospital	Doctor
35	Learner#b864c998-6003-48c7-99f6-9932f95b301b	Nigeria	Graduate Student	University of Ibadan	Pharmacy and Pharmacology
36	Learner#b8a8310a-3d88-40d8-bd65-140da07e6d...	Pakistan	Undergraduate Student	Karachi University	Accounting and Finance
37	Learner#bd7bca97-bbf5-4881-b396-faece268451d	India	NULL	NULL	NULL

- **Flag missing tracking questions Column:**

```
ALTER TABLE public.master_table ADD COLUMN is_tracking_missing BOOLEAN;
UPDATE public.master_table SET is_tracking_missing = tracking_questions IS NULL;
```

Data Output Messages Notifications

```
UPDATE 184710

Query returned successfully in 6 secs 256 msec.
```

- **Tracking Questions Null Preview:**

```
SELECT
  COUNT(*) FILTER (WHERE tracking_questions IS NULL) AS tracking_questions_nulls,
  COUNT(*) FILTER (WHERE apply_date IS NULL) AS apply_date_nulls,
  COUNT(*) FILTER (WHERE institution IS NULL) AS institution_nulls,
  COUNT(*) FILTER (WHERE major IS NULL) AS major_nulls,
  COUNT(*) FILTER (WHERE degree IS NULL) AS degree_nulls,
  COUNT(*) FILTER (WHERE gender IS NULL) AS gender_nulls,
  COUNT(*) FILTER (WHERE assigned_cohort IS NULL) AS assigned_cohort_nulls,
  COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS opportunity_id_nulls,
  COUNT(*) FILTER (WHERE duration_days IS NULL) AS duration_days_nulls,
  COUNT(*) FILTER (WHERE cohort_size IS NULL) AS cohort_size_nulls,
  COUNT(*) FILTER (WHERE end_date IS NULL) AS end_date_nulls
FROM public.master_table;
```

Data Output Messages Notifications

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

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint	cohort_size_nulls bigint	end_date_nulls bigint
1	86471	71296	52707	52716	52707	0	84426	71294	84426	84426	84426

- **Rows with Multiple NULL Columns (Before):**

```sql
SELECT * FROM public.master_table
WHERE opportunity_id IS NULL
  AND assigned_cohort IS NULL
  AND opportunity_name IS NULL
  AND zip IS NULL
  AND city IS NULL
  AND opportunity_code IS NULL;
```

Data Output Messages Notifications

Showing rows: 1 to 1000 Page No: 1 of 43

	ercreatedate text	userlastmodifieddate text	birthdate text	city text	zip text	state text	opportunity_id text	assigned_cohort text	apply_date text	status integer	opportunity_name text	category text	opportunity_code text	tracking_questio text
1	023-07-04T04:48:00.001Z	2023-07-04T04:48:18.344Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
2	024-12-13T06:27:09.183Z	2024-12-13T06:28:00.997Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
3	023-03-15T05:41:41.579Z	2023-03-15T05:42:26.832Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
4	024-08-30T12:51:43.371Z	2024-08-30T12:52:12.943Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
5	024-04-04T12:46:40.055Z	2024-04-04T16:09:03.013Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
6	023-03-11T11:23:08.164Z	2023-03-11T11:23:28.377Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
7	024-10-23T11:01:55.760Z	2024-10-23T11:02:32.067Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]
8	023-06-05T11:57:05.874Z	2023-06-05T11:57:27.905Z	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]	[null]

Total rows: 42599 Query complete 00:00:00.585 CRLF Ln 123,

Deleting all rows from master_table where all of the following columns are NULL at the same time:

- **opportunity_id**
- **assigned_cohort**
- **opportunity_name**
- **zip**
- **city**
- **opportunity_code**

After Deleting the Rows:

```sql
--preview for the table which are nulls together
SELECT * FROM public.master_table
WHERE opportunity_id IS NULL
  AND assigned_cohort IS NULL
  AND opportunity_name IS NULL
  AND zip IS NULL
  AND city IS NULL
  AND opportunity_code IS NULL;
--Deleting the rows which are null together
DELETE FROM public.master_table
WHERE opportunity_id IS NULL
  AND assigned_cohort IS NULL
  AND opportunity_name IS NULL
  AND zip IS NULL
  AND city IS NULL
  AND opportunity_code IS NULL;
```

Data Output Messages Notifications

learner_id text	country text	degree text	institution text	major text	email text	gender text	usercreatedate text	userlastmodifieddate text	birthdate text	city text	zip text	state text	opportunity_id text	assigned_cohort text	apply_date text	status integer	oppor text

Total rows: 0 Query complete 00:00:00.179 CRLF Ln 124, Col 1

- **Count of Nulls After Deleting the Rows**

```sql
-- Count actual NULLs after cleaning
SELECT
  COUNT(*) FILTER (WHERE tracking_questions IS NULL) AS tracking_questions_nulls,
  COUNT(*) FILTER (WHERE apply_date IS NULL) AS apply_date_nulls,
  COUNT(*) FILTER (WHERE institution IS NULL) AS institution_nulls,
  COUNT(*) FILTER (WHERE major IS NULL) AS major_nulls,
  COUNT(*) FILTER (WHERE degree IS NULL) AS degree_nulls,
  COUNT(*) FILTER (WHERE gender IS NULL) AS gender_nulls,
  COUNT(*) FILTER (WHERE assigned_cohort IS NULL) AS assigned_cohort_nulls,
  COUNT(*) FILTER (WHERE opportunity_id IS NULL) AS opportunity_id_nulls,
  COUNT(*) FILTER (WHERE duration_days IS NULL) AS duration_days_nulls
FROM public.master_table;


--gender Update to Unknown for Null Values
```

Data Output | Messages | Notifications

Showing rows: 1 to 1 Page No: 1

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint
1	43872	28697	10431	10440	10431	0	41827	28695	41827

Total rows: 1 Query complete 00:00:00.128

Delete rows with missing duration_days AND no start_date or end_date

The Total null count as shown in the picture below

```sql
--Delete rows with missing duration_days AND no start_date or end_date
DELETE FROM public.master_table
WHERE duration_days IS NULL AND (start_date IS NULL OR end_date IS NULL);
```

Data Output | Messages | Notifications

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

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint	cohort_size_nulls bigint	end_date_nulls bigint
1	14885	2	407	416	407	0	0	0	0	0	0

Updating the Tracking questions to “Not Provided” which resulted in the Shaping the Data

```sql
UPDATE public.master_table
SET tracking_questions = 'Not Provided'
WHERE tracking_questions IS NULL;
```

Data Output | Messages | Notifications

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

	tracking_questions_nulls bigint	apply_date_nulls bigint	institution_nulls bigint	major_nulls bigint	degree_nulls bigint	gender_nulls bigint	assigned_cohort_nulls bigint	opportunity_id_nulls bigint	duration_days_nulls bigint	cohort_size_nulls bigint	end_date_nulls bigint
1	0	2	407	416	407	0	0	0	0	0	0

Total Nulls Percentage After Cleaning the data:

A	B	C	D
	Null Count	Total Rows	Percentage (%)
institution	538	100284	0.54
state	473	100284	0.47
major	419	100284	0.42
zip	412	100284	0.41
city	411	100284	0.41
birthdate	409	100284	0.41
degree	407	100284	0.41
email	84	100284	0.08
usercreate	84	100284	0.08
userlastmo	84	100284	0.08
country	37	100284	0.04
apply_date	2	100284	0

N16

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	learner_id	country_nulls	degree_nulls	institution_nulls	major_null	gender_nulls	email_nulls	city_nulls	zip_nulls	state_nulls	birthdate_nulls	usercreateddat	userlastmodi	apply_date_nulls	opportunity_id_nulls	opportunity_name	category
2	0	0	407	407	416	0	84	409	411	409	409	84	84	2	0	0	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	

Master Table SQL Queries

- https://drive.google.com/file/d/1VDfQ4Vv7hN8co90VrdVciHUhAfRguc6a/view?usp=drive_link

Cleaned Master Table Dataset

- https://drive.google.com/file/d/15QNbwoLhk4K33S3kUzFsSCHj9MGXdvmg/view?usp=drive_link

Data Cleaning Actions for master_table:

- Removed rows missing critical fields (opportunity data + cohort info).
- Replaced NULL gender values with "Unknown".
- Standardized the Regular Null Texts to Actual Nulls,
- Converted timestamp fields and standardized formats.
- Deleted rows with missing duration_days AND no start_date or end_date
- Deleted rows where ALL of the following were missing: opportunity_id, opportunity_name, opportunity_code, assigned_cohort, city, and zip.
- Deleting the Column Tracking Questions which is resulting around 84000 records are empty
- Deleting the rows of null values from Opportunity_id Column

Conclusion Report

The master_table has been thoroughly cleaned, producing a reliable dataset for reporting, modeling, and visualization. Key actions included:

- Removed Incomplete Rows: Deleted records missing critical fields (opportunity_id, opportunity_name, opportunity_code, assigned_cohort, city, zip) and those lacking duration_days, start_date, and end_date.
- Handled Missing Values: Replaced NULL gender with "Unknown" and converted text-based nulls ('null', 'NULL', 'none') to true NULLs.
- Standardized Formats: Converted epoch timestamps to standard formats and enforced consistent title case for degree, institution, and major.
- Eliminated Redundant Data: Dropped the tracking_questions column (~84,000 empty records) and rows with NULL opportunity_id.

Post-cleaning, NULLs are limited to nonessential fields (<0.5% per column), ensuring no impact on analysis. The dataset is now clean, integrated, and ready for actionable insights. Future steps include automated cleaning scripts and periodic quality checks to maintain integrity.