DATA QUALITY REPORT

Team- 1

Team Member Details:

NAME	EMAIL ID
TARINEE	tarinee@s.amity.edu
PRASHANT KASHYAP	prashantkashyap5579@gmail.com
VARUN JOSE MADANU	varunjosemadanu@gmail.com
MUSTAFA OUN	wnvyvcgc@gmail.com
ANKIT RAJ	226301029@gkv.ac.in
ANCHAL KEDIA	anchalkedia45@gmail.com
VAISHNAVI GATTAGONI	gattagonivaishnavi@gmail.com
USWA SHARIQ	uswashariq@gmail.com

Overview:-

After performing ETL operations on six interconnected datasets (User Data, Cohort Data, Opportunity Data, Learner Opportunity Data, Cognito Data), a Master Table was created to enable consolidated analysis. This report validates the structure, integrity, and completeness of the final integrated dataset.

Source of Master Table

The Master Table is the output of a multi-stage ETL process that merged the following datasets:

- **User Data** (user details)
- **Cohort Data** (start/end dates, size)
- **Opportunity Data** (internship details)
- **Learner Opportunity Data** (learner application records)
- **Cognito Data** (user engagement/activity)

SOURCE TABLE	RECORDS
USER DATA	1,29,260
OPPORTUNITY DATA	188
COHORT DATA	640
LEARNER OPPORTUNITY DATA	1,13,603
COGNITO DATA	1,29,179
MASTER TABLE (FINAL)	1,00,201

Data Quality Checks

- **Record Count Validation**

Status: Passed – Master Table contains expected number of records.

- **Duplicate Check**

Query Used:-

```
SELECT learner_id, cd_user_id, opportunity_id, COUNT(*)

FROM mastertable

GROUP BY learner_id,cd_user_id, opportunity_id

 HAVING COUNT(*) > 1;
```

Result: 0 duplicate rows found on composite key.

Status: Passed – No duplicate entries exist.

- **Missing Value Check**

Query Used:-

```
SELECT
  SUM(CASE WHEN user_activity_portal IS NULL THEN 1 ELSE 0 END) AS missing_column1,
  SUM(CASE WHEN assigned_cohort IS NULL THEN 1 ELSE 0 END) AS missing_column2,
  SUM(CASE WHEN cd_user_id IS NULL THEN 1 ELSE 0 END) AS missing_column3,
  SUM(CASE WHEN ud_user_id IS NULL THEN 1 ELSE 0 END) AS missing_column4
```

```
FROM mastertable;
```

Result: 0 NULL rows found.

Handling Approach: Kept as 'NULL' or ‘0’ for visibility;

- **Foreign Key Integrity Validation**

  Purpose: Ensure IDs correctly link across datasets.

  Result: All key relationships (learner_id, opportunity_id, cohort_code,etc) match correctly across joined tables.

- **Data Type Verification**

  Query Used:-

```
SELECT
  pg_typeof(birthdate_clean),
  pg_typeof(apply_date),
  pg_typeof(User_Activity_Portal)
FROM master_table
LIMIT 1;
```

Result: All fields with correct data types: DATE, INTEGER, TEXT .

Issues Identified & Fixes Applied

ISSUE	SQL FIX
null values in multiple columns	Replaced with 'NULL' or ‘0’ using UPDATE
inconsistent foreign key references	Cleaned during join with TRIM, LOWER() functions
non-uniform date formats	Standardized using TO_DATE() and ::DATE
missing cohort start dates from unix timestamps	Converted using: TO_TIMESTAMP(start_date / 1000)::DATE after explicit cast to BIGINT

Final Assessment of Master Table

METRIC	STATUS
record count	Pass
duplicate entries	None
missing values	Handled as 'NULL' or ‘0’
referential links	Intact
data types	Correct
date fields	Uniform (in DATE format)