

# Data Management Governance Bootcamp

## Capstone 2



# The group members 5

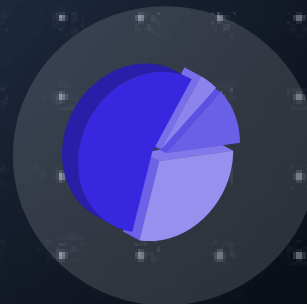


**Sarah Al-Zahrani**



**Seham Salman**

**Maha Sultan**



**Nawal Mohammed**

**Mthayel Almutairi**



# Introduction

## Smart Learning Management Platform

A Learning Management System (LMS) is a comprehensive digital platform designed to streamline the educational process in online environments. It allows easy student enrollment, tracks academic progress, and analyzes learning behavior through detailed reports and data insights. LMS serves as a central tool for enhancing education quality, improving the learner's experience, and supporting data-driven decision-making within educational institutions.

## Project Focus

Enhancing the quality of Learning Management System (LMS) data and analyzing it effectively using advanced cloud-based tools.

The project aims to improve the accuracy and organization of LMS data to ensure its reliability and usability. It also involves leveraging cloud-based analytical tools to extract meaningful insights that support decision-making and enhance the educational process.

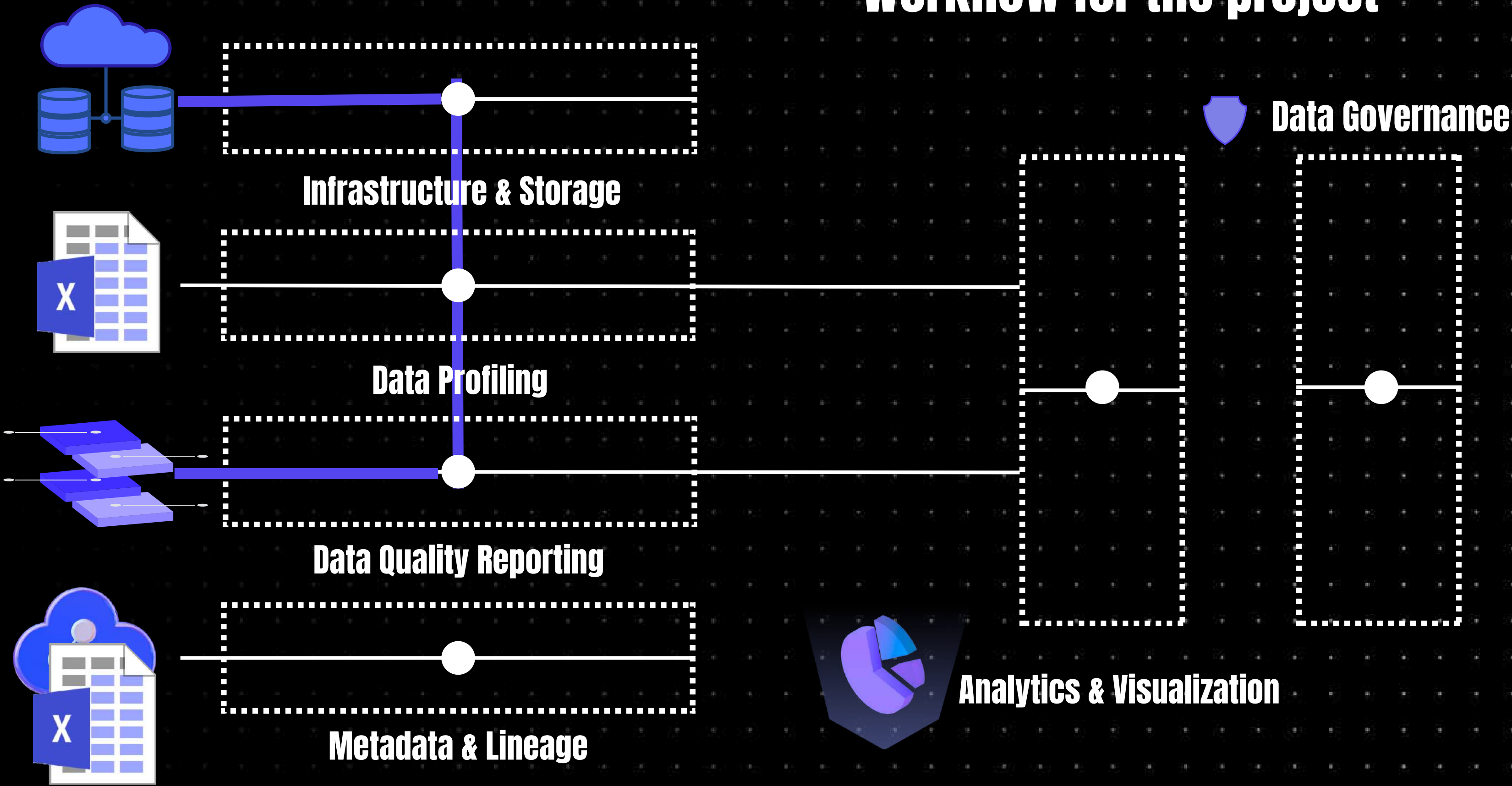
## Data Protection and Compliance

Refers to the protection of sensitive LMS data through the enforcement of governance policies and security controls. It ensures data confidentiality, integrity, and availability while meeting privacy regulations and institutional standards. This helps maintain trust and prevents unauthorized access or misuse of information.

## Data Accuracy Consistency

This refers to ensuring that data is correct, reliable, and uniformly maintained across systems and processes. Accurate data reflects real-world values without errors, while consistent data remains uniform across different platforms and timeframes. Together, they are essential for effective decision-making, reliable reporting, and maintaining the integrity of business and educational systems.

# Workflow for the project





# Data Ingestion and Preparation

- We collected data from three main sources:

- CSV files (containing attendance records)

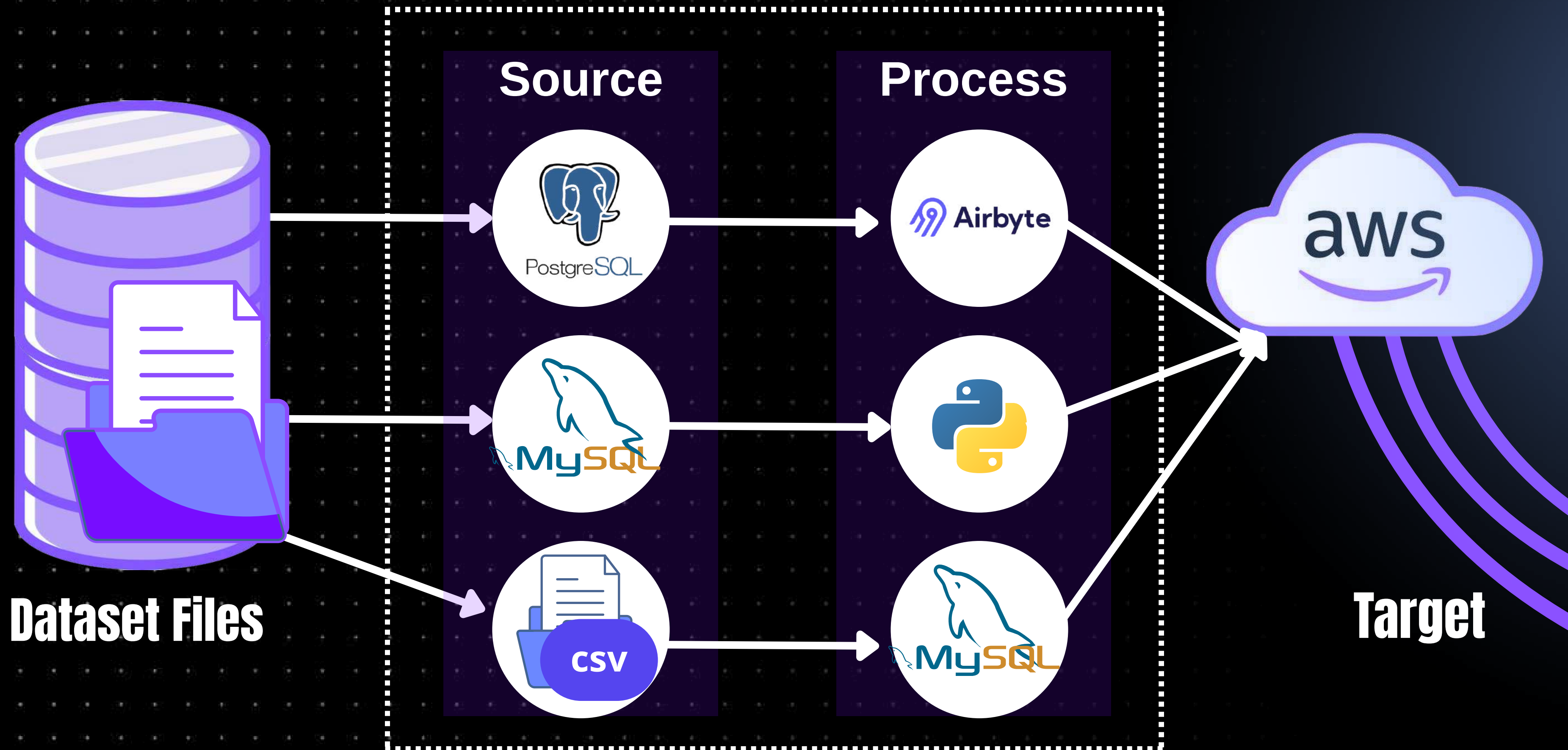
- We built a centralized database using MySQL hosted on AWS RDS.

- Local MySQL database (from Stage 1 of the program)

- PostgreSQL (LMS system database)

- We used Airbyte and SQL and Python to extract and load the data into the RDS database, ensuring schema consistency and valid relationships between tables.

# DATA LINEAGE



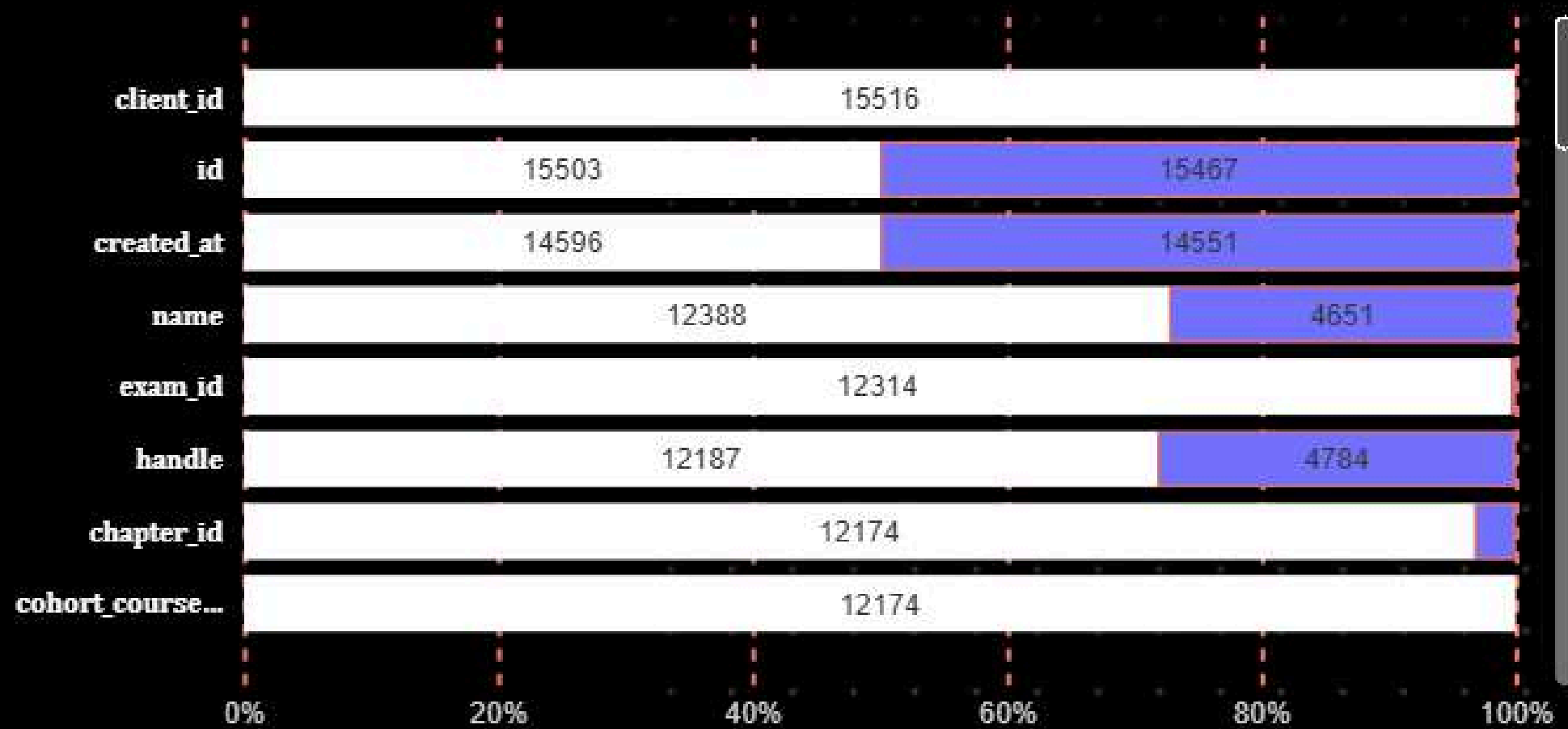
# Data Profiling and Quality Rules

We performed initial profiling using SQL and Python to analyze:

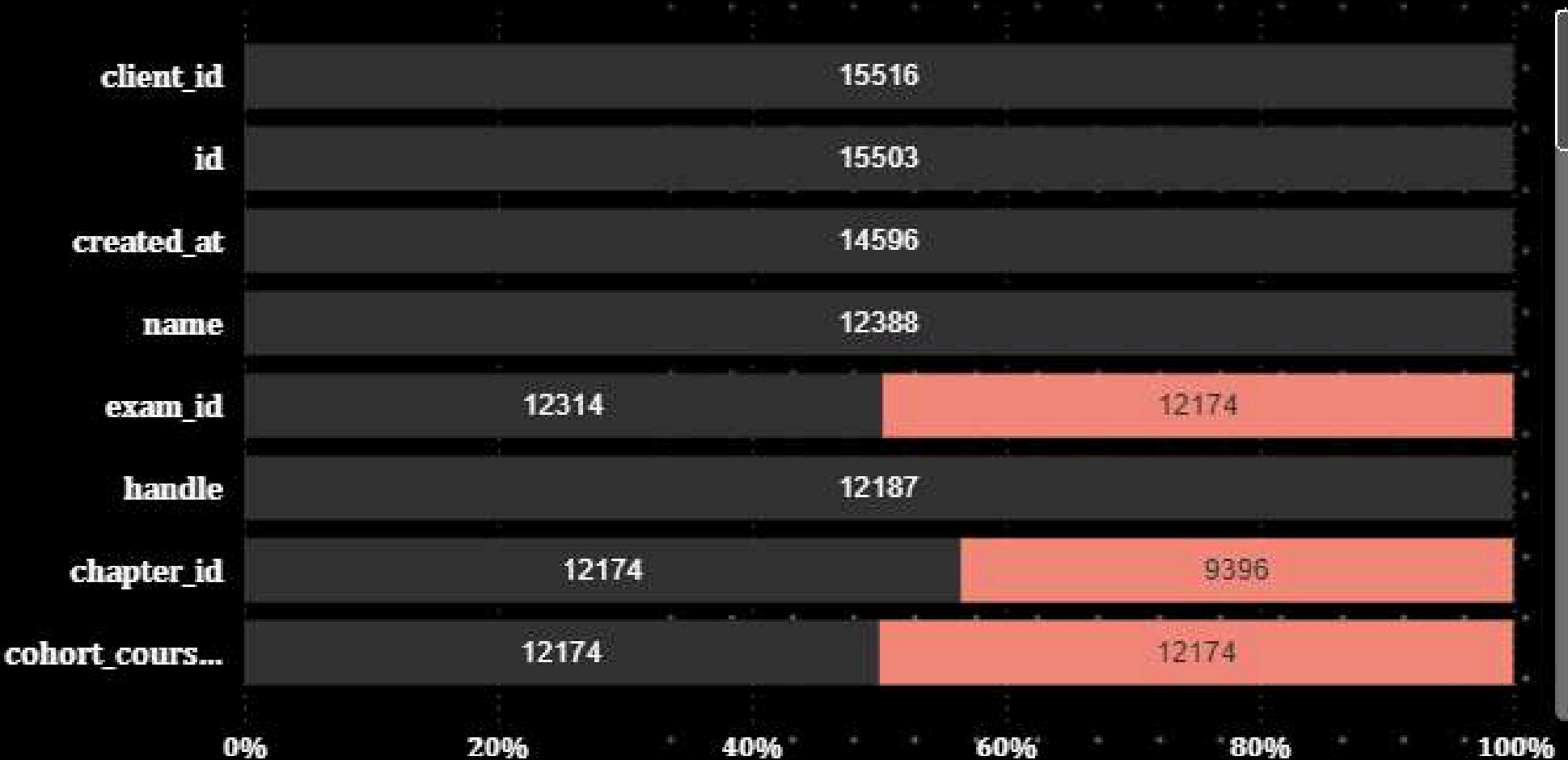
- Missing values (nulls)
- Distinct values
- Field validity (e.g., email format)
- We created Data Quality Rules for critical columns to capture and document inconsistencies.

TableName	Column Name	Data Type	Number of record	Unique Values	Duplicates %	Missing %
cohort enrollments	id	bigint	2208	2208	0	0
	user_id	bigint		856	61.23%	0
	client_id	text		2	99.91%	0

● Number of Records ● Unique of Records



● Number of Records ● Missing of Records



## Dashboard for data quality procedure analysis

Uniqueness

46K

Rows Processed


192K

Completeness

135K



# Rules

Num	CriticalData Element (CDE)	 Rule
1	score	score must be between 0 and 100.
2	question_ids	should be formatted correctly, e.g., separated by ---\n-
3	name	[Track   Domain] - [Start Month] - [Attendanc Type : Online / Onsite] - [Learning Type : Self Paced / Instructor led].
4	filename	[project_or_model][task_type][topic]_[optional_tag].[extension]
5	name	[kind] -[Topic / Title] - [Date (optional) ]

# Observation & Recommendations

Observation table	Column Name	Observation
Cohorts	days	Some entries are empty rather than null. These fields should be filled with valid day values (e.g., {Mon, Tue, Wed, Thu, Fri}).
Cohorts	name	Some entries do not follow the standard format (e.g., "M2.3-GenAI-Intro-B2").
Cohorts	times	Some entries are empty rather than null. These fields should be filled with valid time values (e.g., "09:00-10:00", "14:00-15:30")
handin attachments	filename	Ensure that all filenames adhere to the expected format, such as "Assignment_Code.zip"
programs	description	HTML code is present (e.g., <div><h1>GenAI Introduction</h1></div>). Remove HTML and use plain text.

# Data Cleaning and Transformation

- Merging some files like attendance
- Removing duplicates using **DISTINCT**
- Handling missing values by filling it
- Standardizing date formats with **DATE\_FORMAT**
- Trimming extra spaces using **TRIM**
- Merging datasets using **UNION**
- Saving the cleaned data into new tables in the **lms\_dev** schema

# Data Governance and PDPL Compliance

## Data Governance Principles:

**Data Quality:** Regular profiling and validation ensure accurate, complete, and consistent data.  
**Data Integrity:** Controlled access ensures that only authorized users can modify data.

**Data Transparency:** Clear data documentation with metadata for all tables and columns.



# **Data Governance and PDPL Compliance**

## **PDPL Compliance:**

### **Data Privacy and Masking:**

---

**Sensitive information (student emails, first names, and last names) is masked to protect personal data.**

### **User Rights Management:**

---

**Users can access, correct, or delete their data upon request.**

### **Data Security:**

---

**access control are applied to ensure data protection.**

### **Data Minimization:**

---

**Only necessary data is collected and stored, and outdated data is removed.**

### **Compliance Monitoring:**

---

**Regular audits ensure adherence to PDPL requirements.**



# Data Privacy and Masking

id	role	email	last_name	first_name
9405	{student}	Stephen.Vaughn@gmail.com	Vaughn	Stephen
10089	{student}	Marc.Kane@gmail.com	Kane	Marc
8791	{student}	Scott.Gonzalez@gmail.com	Gonzalez	Scott

id	role	email	last_name	first_name
9405	{student}	S*****@gmail.com	V*****	S*****
10089	{student}	M*****@gmail.com	K***	M***
8791	{student}	S*****@gmail.com	G*****	S****

**Masking  
Sensitive  
information**

# Metadata Report

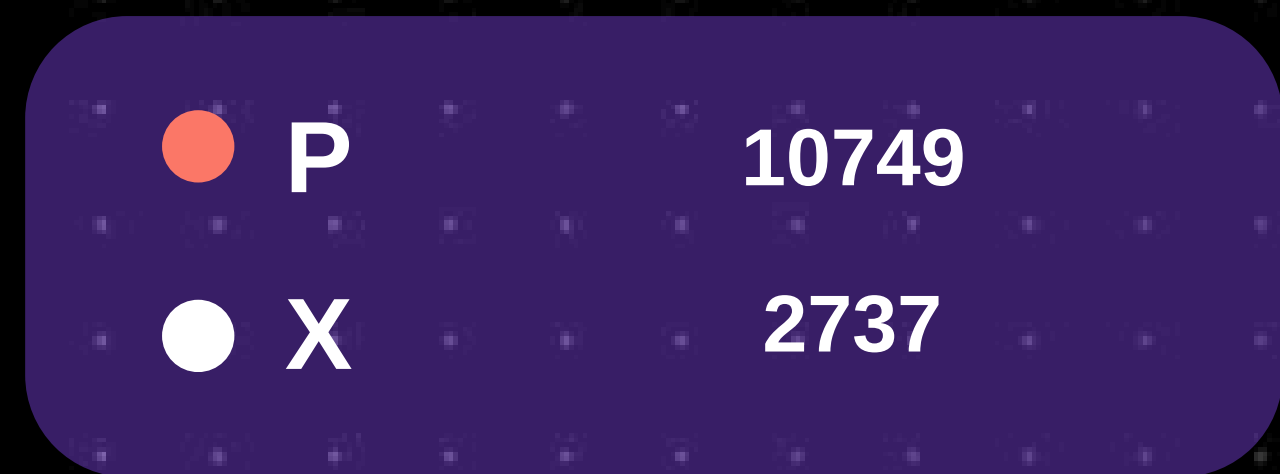
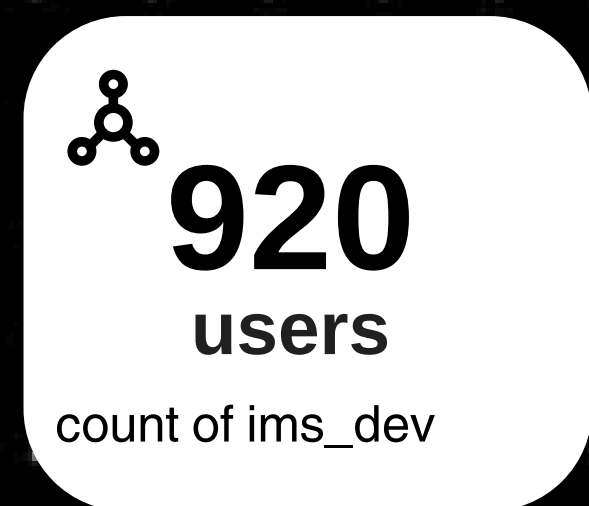
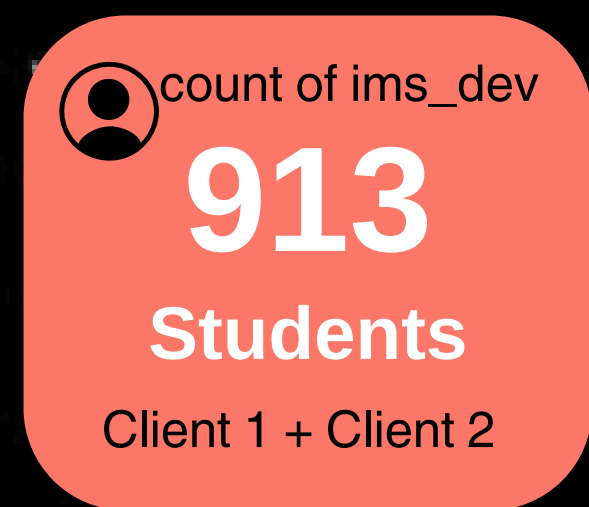
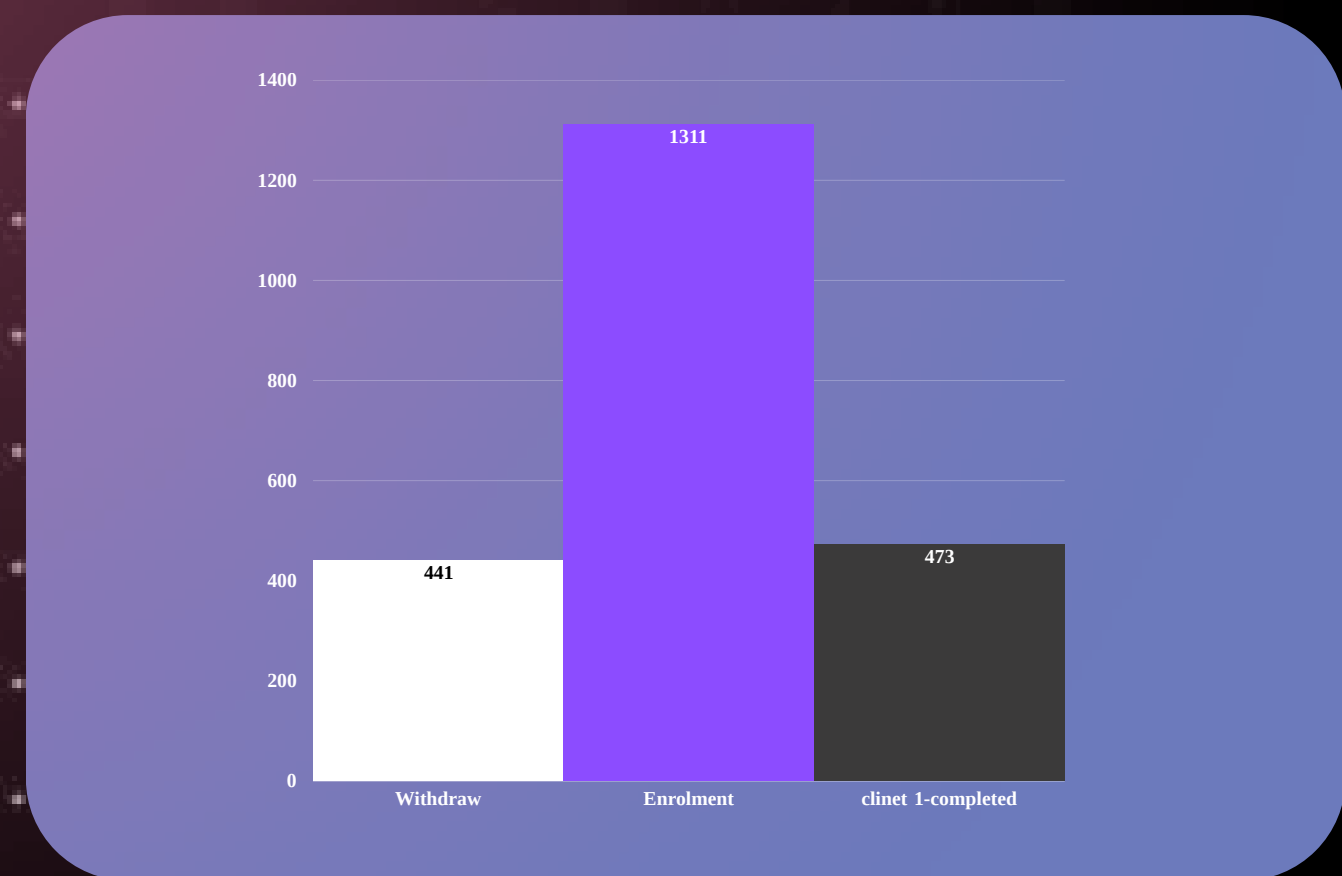
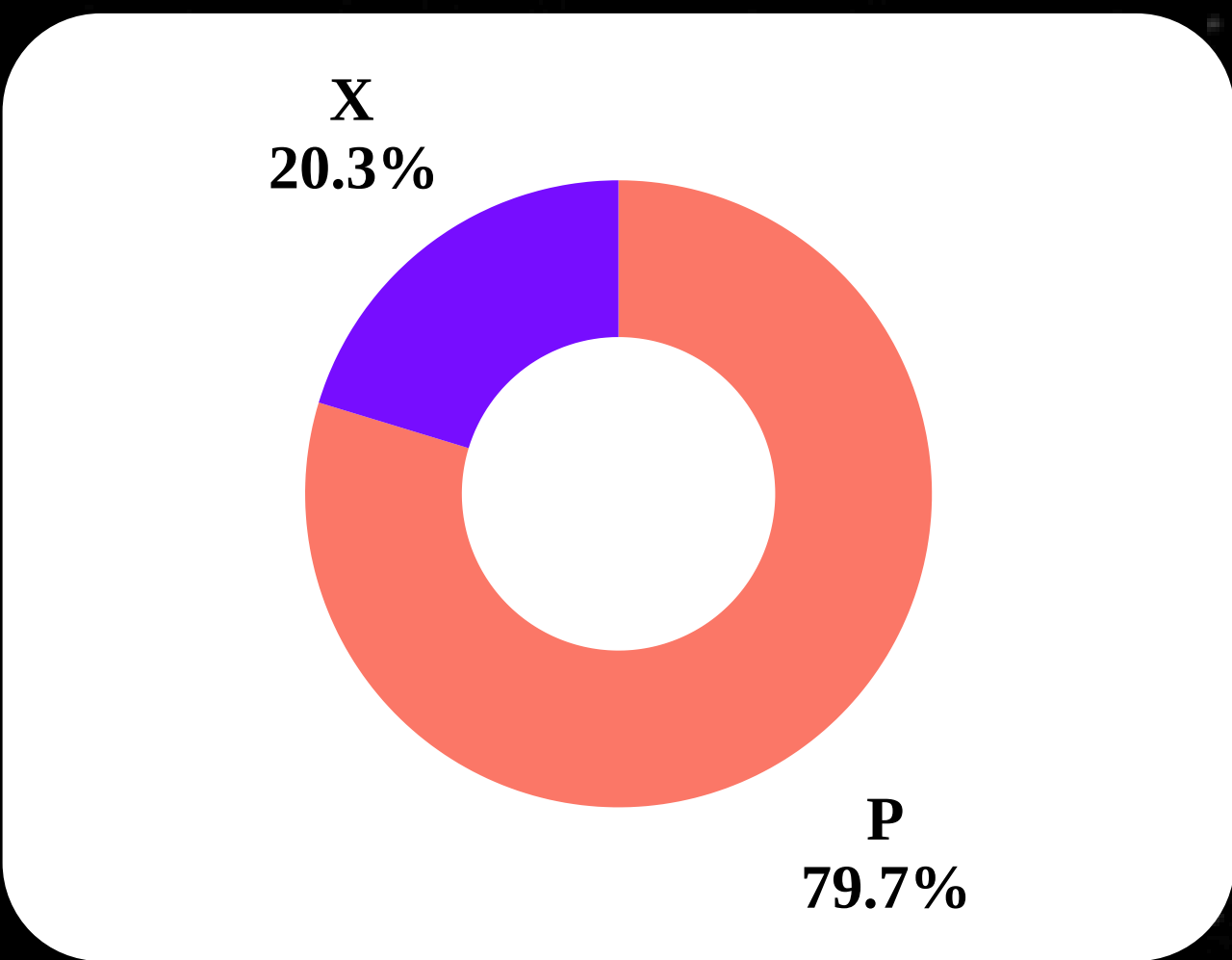
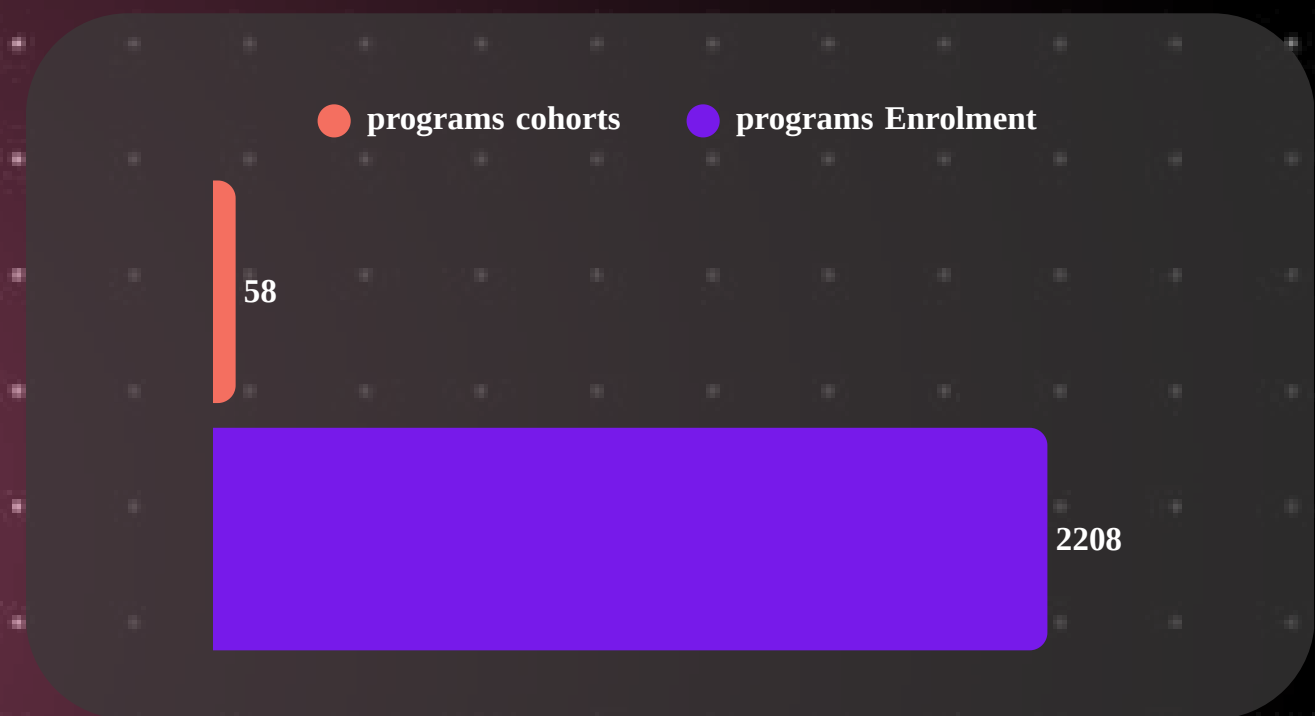


DataSet	Source	Process	Target	Attributes	Data Type
attendance	Bootcamp Attendance Records CSVs and local database SQL Script	SQL Python	AWS RDS SQL	Student_id	text
				Course_Name	text
				Cohort	text

Attributes describe	Master and Reference	Masking
Unique identifier for each student	Master Data	All the column No Need
Name of the course the student is enrolled in	Master Data	
Group or batch to which the student belongs	Master Data	

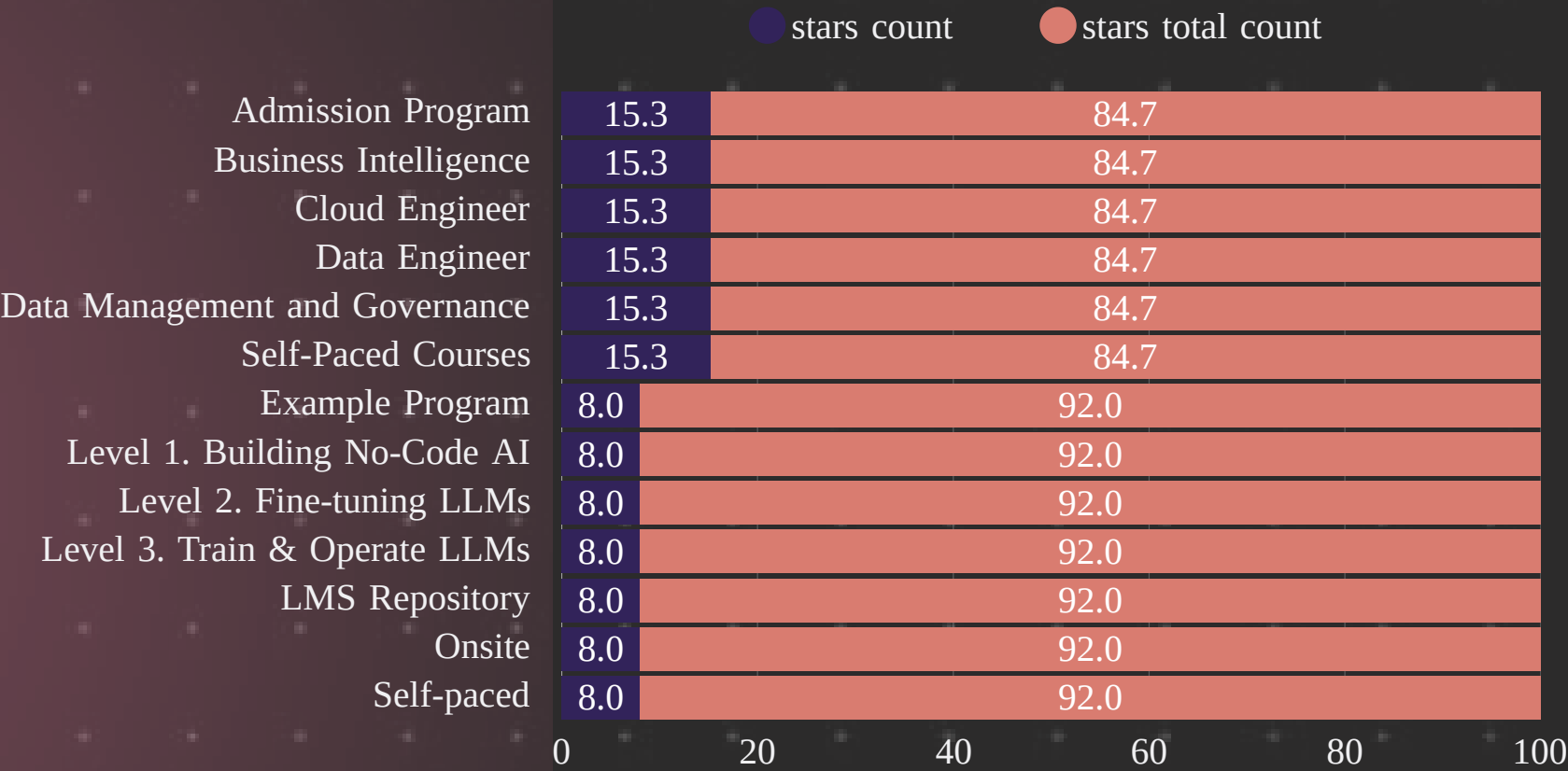


# Overview Business requirements

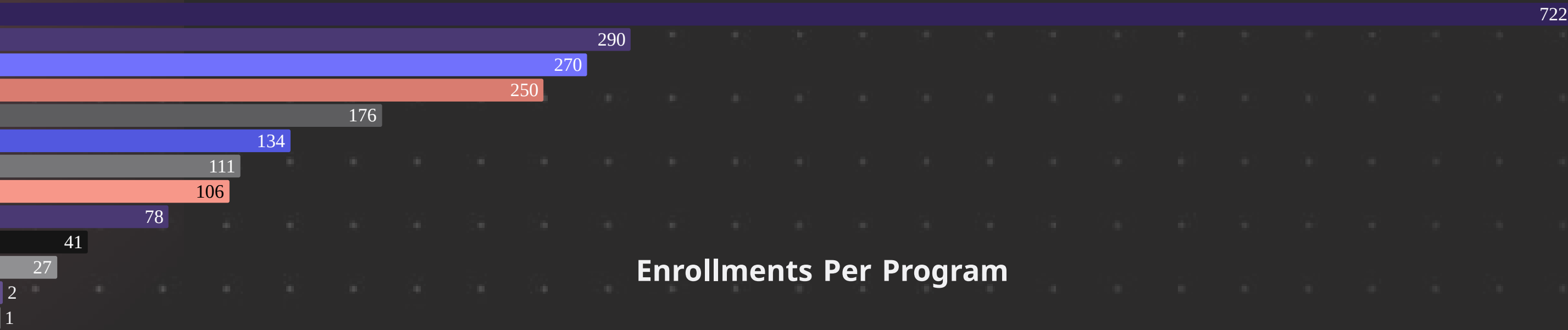
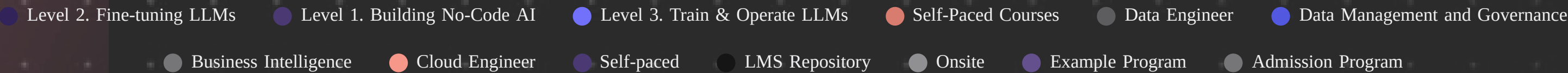
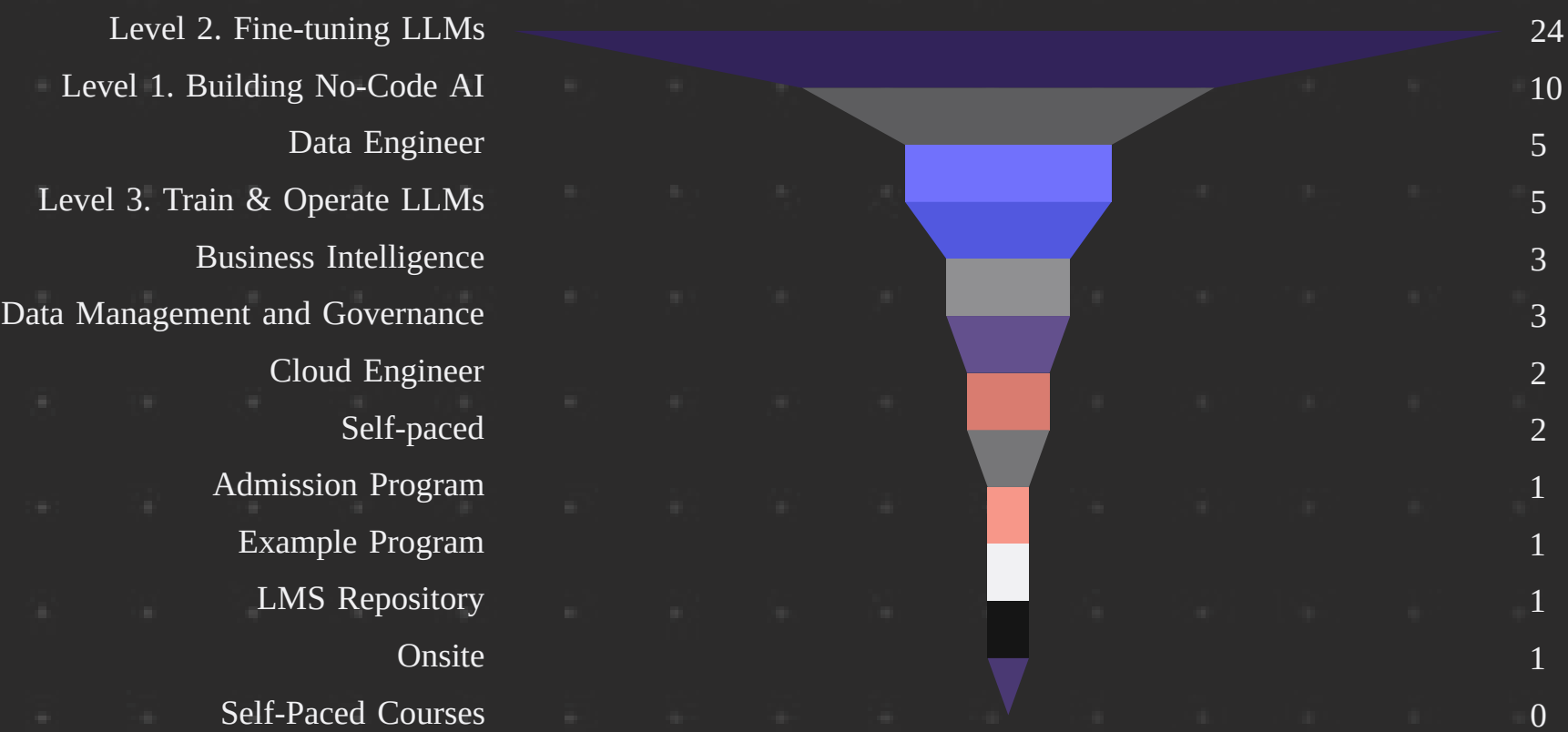


# dashboard

Average Progress by Program



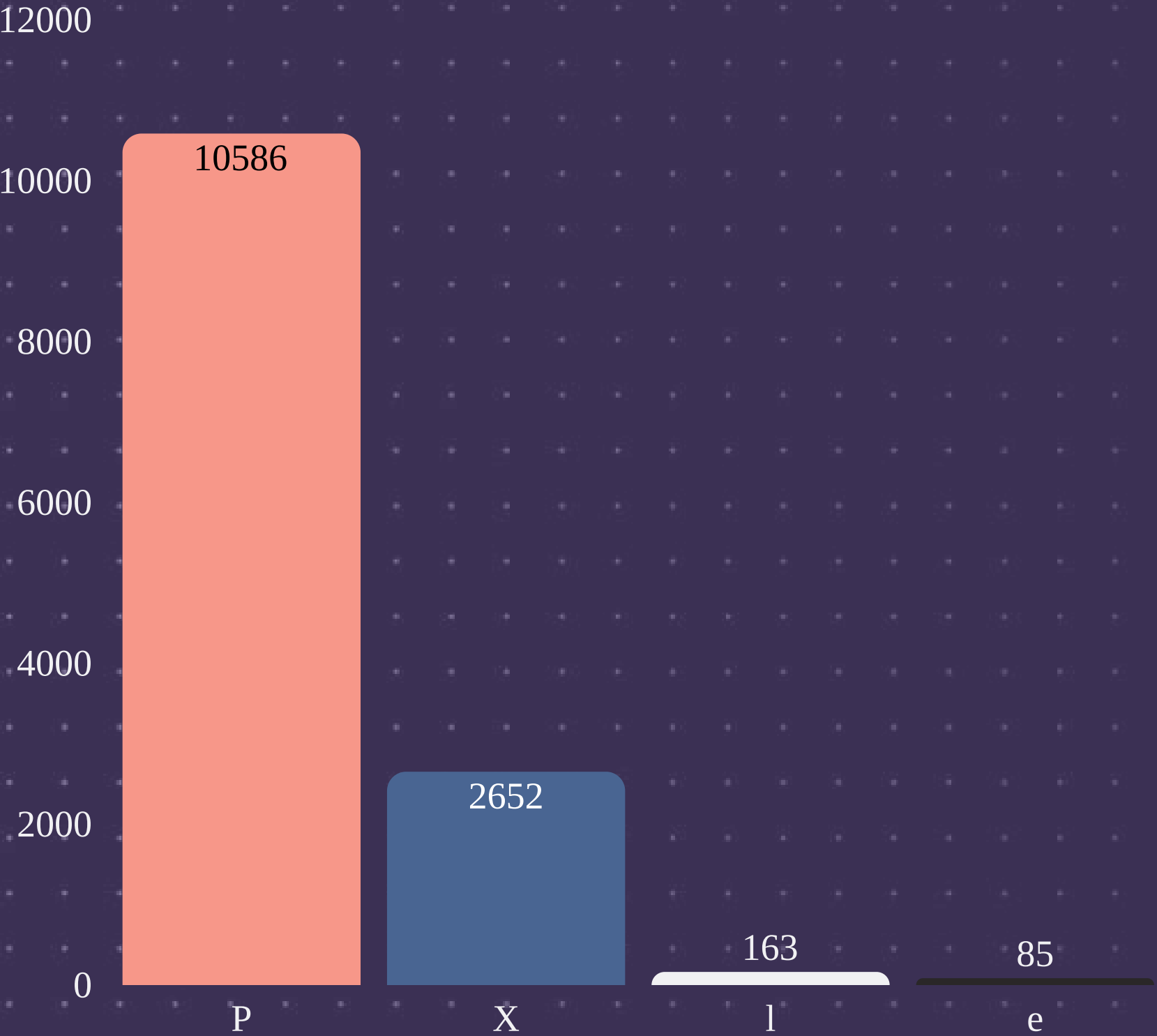
Number of cohorts in each Programs



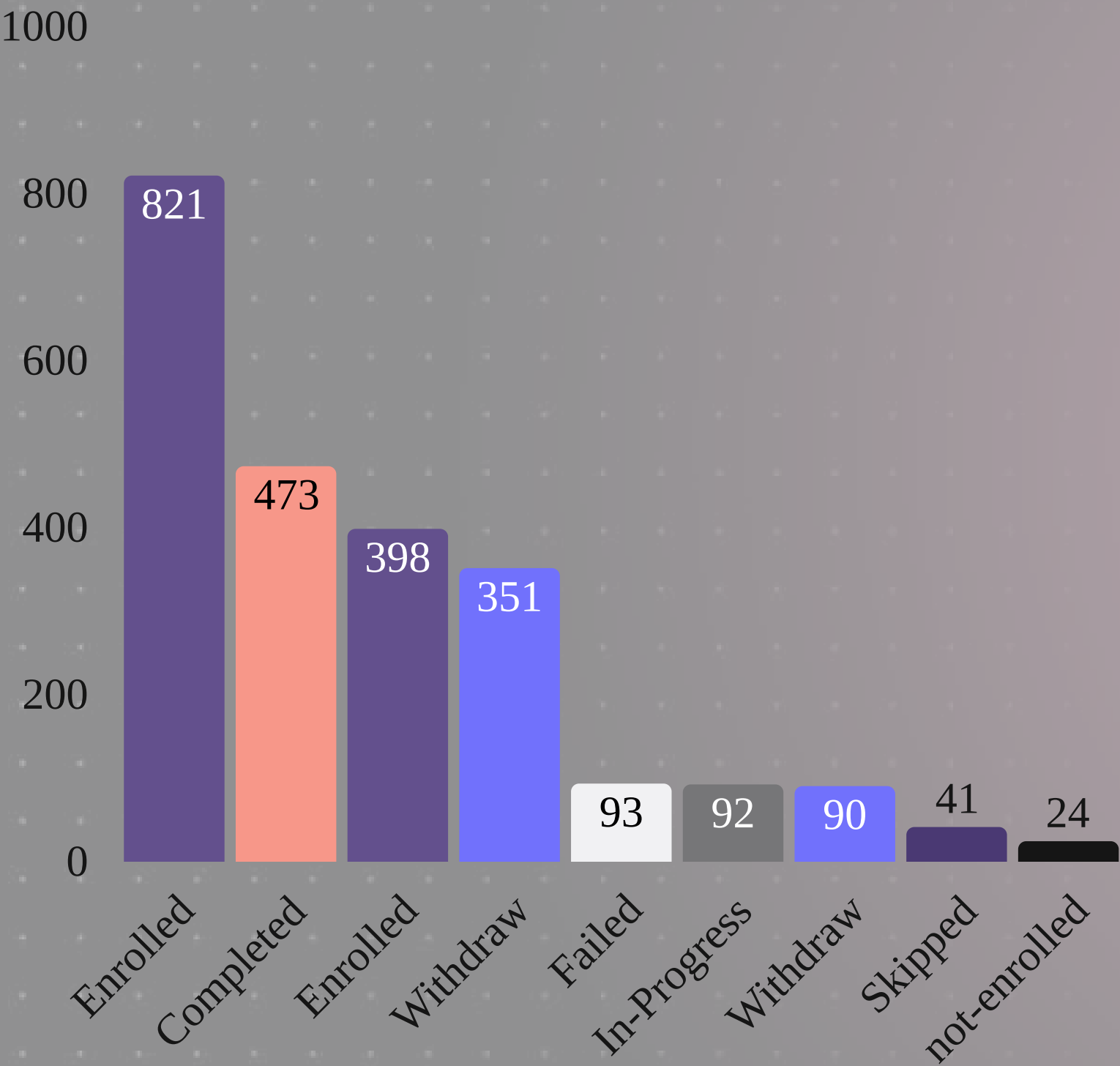
Enrollments Per Program

# dashboard

Overall Attendance Distribution



Student Statuse Distribution





# Challenges Faced in the Project:

## Data Integration from Multiple Sources

We worked with datasets from three different sources and loading into a single RDS database.

1

## Merging Attendance Records

Attendance files for Client 1 and Client 2 were separate. We standardized their formats and merged them into one file

2

## Inconsistent Student Status Across Clients

Student statuses (e.g., Enrolled, Withdrawn) were recorded differently across multiple tables for each client

3

# Conclusion



**Thanks!**