Aakriti Lnu (al1745), Gokula Ranga Naveen Chapala (gc3522),
Muhammad Raees (mr2714), Prajjwal Mehta (pm8607)

This report explains the process of creating a relational database in Phase I of the project. The dataset we selected is Meta Kaggle (https://www.kaggle.com/datasets/kaggle/meta-kaggle). The report briefly describes the dataset, a basic relational model, and details about the program that loads a subset of the dataset into the database. The dataset is very large, therefore, we selected a subset of it having more than 127 million tuples.

Note: The program code is provided in the zipped folder. The program can be executed by following the "README.md" instructions. The global settings are provided in "globals.py" inside the "funcs" subfolder. The program to clean files ("clean_files.py") takes around 100 seconds to filter a few input data files provided in the data directory ("globals.py"). The program to load the dataset into the database ("app.py") takes around 80 minutes to run.

Meta Kaggle Dataset Overview

The "Meta Kaggle" dataset provides a comprehensive view of the Kaggle community and platform activities, capturing various entities such as users, competitions, datasets, and kernels (code notebooks) and how they interact with one another. It enables the exploration of trends in user engagement, competition participation, dataset usage, and shared content within the community. This extensive dataset offers valuable metadata across different entities, facilitating deep analysis of Kaggle's operations and user behaviors, ultimately uncovering insights into how the platform functions as a thriving data science community and competition hub.

Key Insights the Dataset Offers

- **User Engagement**: You can analyze user growth, competition participation, and notebook creation trends.
- **Competitions**: You can explore the number of competitions hosted, the types of competitions that attract the most participants, and how rewards correlate with participation levels.
- **Submissions**: It allows you to track submissions made during competitions, evaluate their timing, and correlate them with user or team performance.
- **Kernels (Notebooks)**: You can analyze the popularity of notebooks in terms of upvotes, trends in coding languages used, and the expertise of notebook creators.
- **Datasets**: Trends in dataset creation and usage can be explored to understand which datasets are most valuable to the community.
- **Discussions**: Forum posts and their interactions provide insights into active discussions and topics of interest within the Kaggle community.

In the following table, we provide a brief overview of selected tables (a subset of Kaggle Meta) and contained records after inserting them into the database. Sections following the table explain the attributes analyzed for each table.

Table Name	Description	Total Records
Users	Contains information about Kaggle users, including their performance tiers and registration data.	20485253
Tags	Contains information about tags applied to competitions, datasets, kernels, and forums.	821
Forums	Captures information about forum posts, including title and parent relationships.	421293
Organizations	Stores details of organizations on Kaggle, including creation date and descriptions.	1601
UserOrganizations	Links users with their affiliated organizations on Kaggle.	2864
UserFollowers	Tracks who follows whom among Kaggle users.	1525039
Datasets	Stores dataset information, including total downloads, views, and votes.	388889
DatasetTags	Associates datasets with tags for easier categorization and search.	358480
Competitions	Contains details about competitions, including deadlines, rewards, and evaluation methods.	5695
CompetitionTags	Links competitions to relevant tags for categorization purposes.	1046
Teams	Contains team information for competitions, such as membership and medals won.	7675351
Submissions	Tracks submissions to competitions, including scores and submission dates.	14800846
UserAchievements	Records achievements of Kaggle users, including rankings, points, and medal counts.	81940704
	Total	127607882

Dataset Description

To reduce, the complexity of handling uncleaned values, we excluded adding some attributes to the database. The file "clean_files.py" contains the code to remove some data attributes. We deduce data types and extract null counts by analyzing each original file using the code provided in the "analyze_data.py" file. The following sections explain the attributes for each selected table and the set of original attributes it contains.

Users - This table contains information about the users of Kaggle. It allows you to explore user demographics, joining trends, and the various roles that users play on the platform.

Field Name	Description	Data Type	Null Count
Id	Unique identifier for each user. **(PK)**	Integer	0
Country	The country where the user is located.	Text	19034320
DisplayName	The display name of the user.	Text	314
PerformanceTier	User's performance tier on Kaggle.	Integer	0
RegisterDate	The date the user registered on Kaggle.	Date/Time	0
UserName	Username chosen by the user.	Text	1

Tags - The table contains information about tags used on Kaggle for various purposes, such as categorizing datasets, competitions, or notebooks. Tags serve as metadata to help users discover relevant content. It can help identify the most popular topics on Kaggle and which areas of data science are frequently discussed or worked on.

Field Name	Description	Data Type	Null Count
Id	Unique identifier for the tag. **(PK)**	Integer	0
CompetitionCount	Number of competitions associated with this tag.	Integer	0
DatasetCount	Number of datasets associated with this tag.	Integer	0
Description	Textual field providing a detailed explanation or summary of the tag's meaning or purpose.	Text	675
FullPath	This field represents the full hierarchical path to the tag, starting from the top-level parent tag and going down through any sub-tags.	Text	0
KernelCount	Number of kernels (notebooks) associated with the tag.	Integer	0
Name	Name of the tag (e.g., Python, Machine Learning).	Text	0
ParentTagId	ID of the parent tag if applicable, defining tag hierarchy.	Integer	10
Slug	URL-friendly version of the tag name. Often used in web applications for clean URLs.	Text	0

Forums - This table holds information about forum posts made on Kaggle. Forums provide a platform for users to ask questions, share knowledge, and collaborate on projects.

Field Name	Description	Data Type	Null Count
Id	Unique identifier for the forum post. **(PK)**	Integer	0
ParentForumId	ID of the parent forum post if it is a reply.	Integer	27
Title	Title of the forum post.	Text	189

Organizations - This table contains information about organizations associated with Kaggle. This helps in understanding which organizations are most active or have a significant presence in Kaggle competitions and contributions.

Field Name	Description	Data Type	Null Count
Id	Unique identifier for the organization. **(PK)**	Integer	0
CreationDate	The date when the organization was added.	Date/Time	0
Description	Detailed description of the organization.	Text	1156
Name	Name of the organization.	Text	0
Slug	URL-friendly version of the organization name.	Text	0

UserOrganizations - This table links users with their organizations. This table provides insights into user affiliations with different companies or institutions, helping analyze which organizations have the most active Kaggle users.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for each user-organization relationship. **(PK)**	Integer	0
JoinDate	The date when the user joined the organization.	Date/Time	0
OrganizationId	ID of the organization the user is affiliated with. **(FK)**	Integer	0
UserId	ID of the user associated with the organization. **(FK)**	Integer	0

UserFollowers - This table provides data about relationships between users, specifically, who is following whom. This table allows us to analyze the follower-followee relationships within the Kaggle community, helping us understand influential users and community networking.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for each follower relationship. **(PK)**	Integer	0
CreationDate	The date when the follower relationship was created.	DateTime	0
FollowingUserId	The ID of the user following another user. **(FK)**	Integer	0
UserId	The ID of the user being followed. **(FK)**	Integer	0

Datasets - This table contains information about datasets uploaded to Kaggle.This table is important for understanding dataset contributions and their popularity in the Kaggle community.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the dataset. **(PK)**	Integer	0
CreationDate	The date the dataset was created.	Date/Time	0
CreatorUserId	ID of the user who created the dataset. **(FK)**	Integer	0
CurrentDatasetVersionId	ID of the current version of the dataset.	Integer	176
CurrentDatasourceVersionId	ID of the current version of the data source, if applicable.	Integer	187
ForumId	ID of the forum associated with the dataset. **(FK)**	Integer	0
LastActivityDate	Date of the last activity related to the dataset.	Date/Time	0
OwnerUserId	ID of the user who owns the dataset.	Integer	2366
TotalDownloads	Total number of times the dataset has been downloaded.	Integer	0
TotalKernels	Total number of kernels (notebooks) created using the dataset.	Integer	0
TotalViews	Total number of views the dataset has received.	Integer	0
TotalVotes	Total number of votes (upvotes) the dataset has received.	Integer	0

DatasetTags - This table links datasets to their associated tags. This table allows us to explore the categorization of datasets, identify trends in popular data topics, and analyze which datasets are frequently used by the Kaggle community.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the dataset-tag relationship. **(PK)**	Integer	0
DatasetId	The ID of the dataset associated with the tag. **(FK)**	Integer	0
TagId	The ID of the tag associated with the dataset. **(FK)**	Integer	0

Competitions - The table provides essential information about the setup, rules, and logistics of Kaggle competitions. This structure allows Kaggle to manage various types of competitions, track participant activities, and evaluate performance metrics in a standardized way. This table stores details about Kaggle competitions, including their names, descriptions, start and end dates, and host organizations.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the competition. **(PK)**	Integer	0
Slug	URL-friendly version of the competition title.	Text	0
Title	Name of the competition.	Text	0
Subtitle	Subtitle or tagline of the competition.	Text	8
HostSegmentTitle	Category or type of host (e.g., Getting Started, Research).	Text	0
ForumId	ID of the forum associated with the competition. **(FK)**	Integer	0
OrganizationId	ID of the organization hosting the competition.	Integer	5194
EnabledDate	The date when the competition was opened.	Date/Time	0
DeadlineDate	The final date for submissions.	Date/Time	0
ProhibitNewEntrantsDeadlineDate	Date after which no new entrants are allowed.	Date/Time	5332
TeamMergerDeadlineDate	Date by which teams can merge.	Date/Time	5338
TeamModelDeadlineDate	Final date for team model submissions.	Date/Time	5406
ModelSubmissionDeadlineDate	Deadline for model submissions.	Date/Time	5679
FinalLeaderboardHasBeenVerified	Boolean indicating if the final leaderboard has been verified.	Boolean	0
HasKernels	Boolean indicating if kernels are allowed.	Boolean	0
OnlyAllowKernelSubmissions	Boolean indicating if only kernel-based submissions are allowed.	Boolean	0
HasLeaderboard	Boolean indicating if a public leaderboard is available.	Boolean	0
LeaderboardPercentage	Percentage of the test set used for the public leaderboard.	Float	0
ScoreTruncationNumDecimals	The number of decimal places scores are truncated to.	Integer	0
EvaluationAlgorithmAbbreviation	The short version of the evaluation algorithm is used.	Text	1
EvaluationAlgorithmName	Full name of the evaluation algorithm used.	Text	1
EvaluationAlgorithmDescription	A detailed description of how the evaluation algorithm works.	Text	26

EvaluationAlgorithmIsMax	Boolean indicating if higher scores are better.	Boolean	1
MaxDailySubmissions	Maximum number of submissions allowed per day.	Integer	0
NumScoredSubmissions	Number of scored submissions displayed on the leaderboard.	Integer	0
MaxTeamSize	Maximum team size allowed in the competition.	Integer	0
BanTeamMergers	Boolean indicating if team mergers are prohibited.	Boolean	0
EnableTeamModels	Boolean indicating if team models are allowed.	Boolean	4774
RewardType	Type of reward (e.g., monetary, job opportunity).	Text	4280
RewardQuantity	The total reward amount for the competition.	Float	0
NumPrizes	Number of prizes offered in the competition.	Integer	0
UserRankMultiplier	The multiplier is applied to user rankings.	Float	0
CanQualifyTiers	Boolean indicating if participation qualifies users for tiers.	Boolean	694
TotalTeams	Total number of teams participating.	Integer	719
TotalCompetitors	Total number of individual participants.	Integer	1475
TotalSubmissions	Total number of submissions made.	Text	5693
ValidationSetName	The name of the validation set.	Float	5693
ValidationSetValue	The value or proportion of the validation set used.	Boolean	0
EnableSubmissionModelHashes	Boolean indicating if submission model hashes are enabled.	Boolean	0
EnableSubmissionModelAttachments	Boolean indicating if submission model attachments are enabled.	Text	5693
HostName	Name of the competition host.	Integer	0
CompetitionTypeId	Unique identifier for the type of competition (e.g., featured, research).	Integer	0

CompetitionTags - This table links competitions to their associated tags, categorizing competitions into relevant topics. This table helps us understand the distribution of topics across competitions and which tags (topics) are most frequently applied to competitions.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the competition-tag relationship. **(PK)**	Integer	0
CompetitionId	The ID of the competition is associated with the tag. **(FK)**	Integer	0
TagId	ID of the tag associated with the competition. **(FK)**	Integer	0

Teams - This table contains information about teams participating in Kaggle competitions. This table is critical for analyzing team-based competition behavior and performance in Kaggle competitions.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the team. **(PK)**	Integer	0
CompetitionId	The ID of the competition the team participated in. **(FK)**	Integer	0

IsBenchmark	Boolean indicates if the team is a benchmark team.	Boolean	0
LastSubmissionDate	The date of the team's last submission.	Date/Time	6833008
Medal	The medal earned by the team, if applicable.	Text	7589049
TeamLeaderId	ID of the team leader.	Integer	26306
TeamName	The name of the team.	Text	845

Submissions - This table tracks submissions to Kaggle competitions. Analyzing this table helps track competition dynamics and performance based on the timing and quality of submissions.

Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for the submission. **(PK)**	Integer	0
IsAfterDeadline	Boolean indicating if the submission was made after the deadline.	Boolean	0
PrivateScoreFullPrecision	Private score of the submission with full precision.	Float	579285
PrivateScoreLeaderboardDisplay	Private score as displayed on the leaderboard.	Float	579285
PublicScoreFullPrecision	Public score of the submission with full precision.	Float	579285
PublicScoreLeaderboardDisplay	Public score as displayed on the leaderboard.	Float	579097
SubmissionDate	The date the submission was made.	Date/Time	0
SubmittedUserId	ID of the user who submitted the entry.	Integer	1346
TeamId	ID of the team that made the submission. **(FK)**	Integer	0

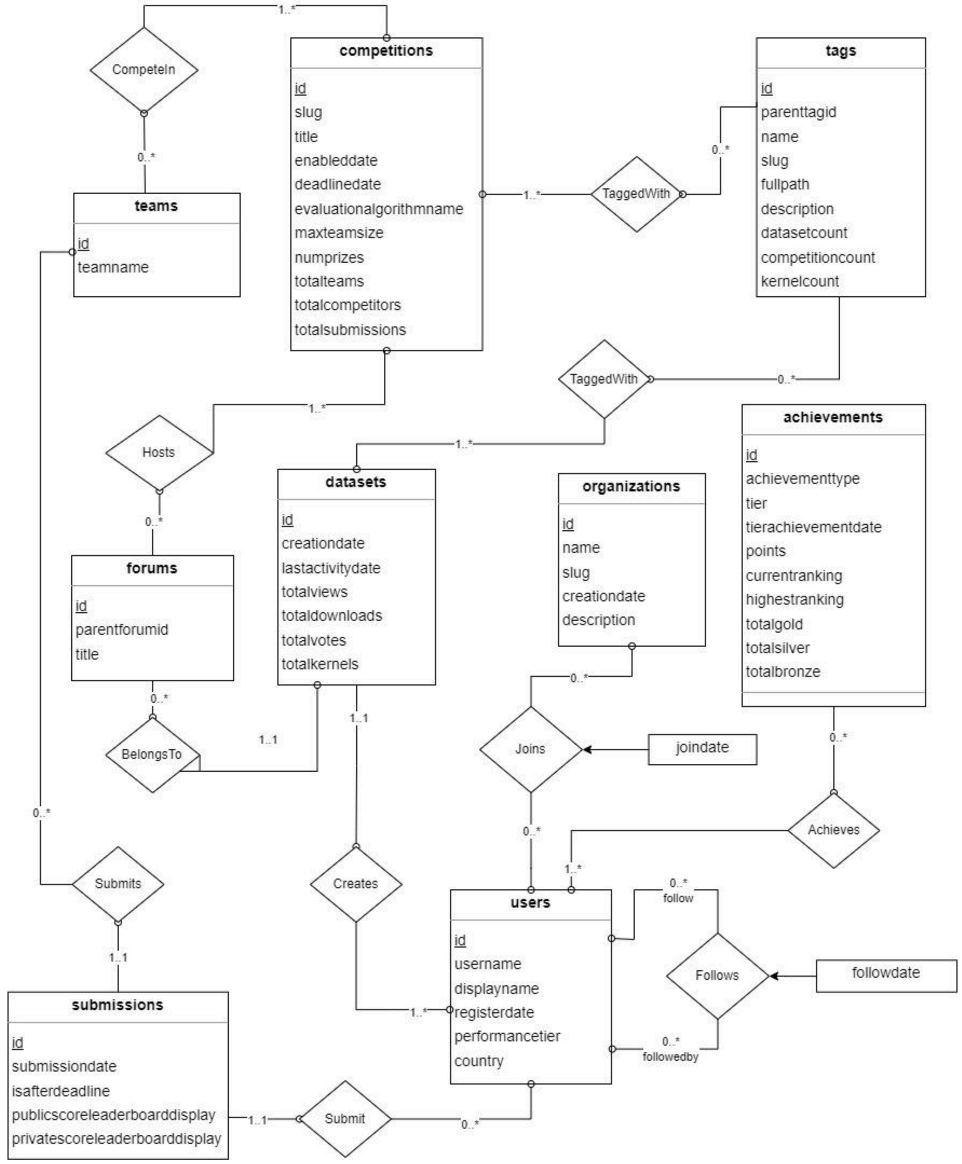
UserAchievements - This table captures the achievements earned by users on the platform. Achievements can be badges, titles, or other recognition for completing specific tasks. Analyzing this table reveals what achievements are most common and how users progress in terms of accolades on Kaggle.

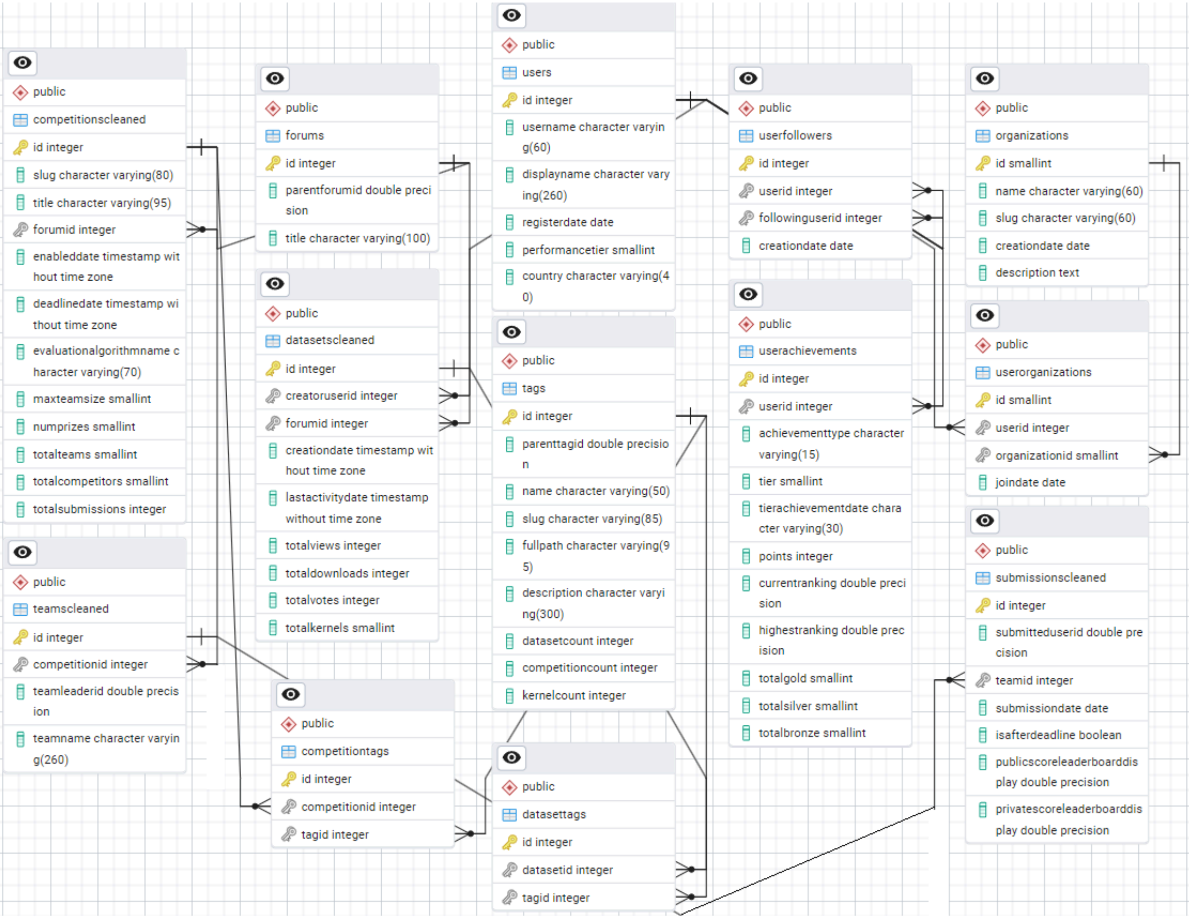
Field Name	**Description**	**Data Type**	**Null Count**
Id	Unique identifier for each achievement. **(PK)**	Integer	0
AchievementType	Type of achievement (e.g., Competitions, Datasets).	Text	0
CurrentRanking	The current global rank of the user on Kaggle is based on the total points they have earned.	Integer	81417741
HighestRanking	The highest global rank the user has ever achieved.	Integer	81417236
Points	Total points earned by the user in Kaggle competitions, kernels, datasets, or discussions.	Integer	0
Tier	User’s performance level or rank category (e.g., 0 - Novice, 1 - Expert) within Kaggle	Integer	0

TierAchievementDate	Date when the user achieved the tier.	Date/Time	494184
TotalBronze	Total number of bronze medals earned by the user.	Integer	0
TotalGold	Total number of gold medals earned by the user.	Integer	0
TotalSilver	Total number of silver medals earned by the user.	Integer	0
UserId	The ID of the user associated with the achievement. **(FK)**	Integer	0

ER and Relational Model

The following diagram shows basic entity-relationship and relational models on the subset of data attributes on the cleaned (and selected) dataset. The given entity sets faithfully represent the dataset.




public
competitionscleaned
id integer
slug character varying(80)
title character varying(95)
forumid integer
enableddate timestamp without time zone
deadlinedate timestamp without time zone
evaluationalgorithmname character varying(70)
maxteamsize smallint
numprizes smallint
totalteams smallint
totalcompetitors smallint
totalsubmissions integer
public
teamscleaned
id integer
competitionid integer
teamleaderid double precision
teamname character varying(260)
public
forums
id integer
parentforumid double precision
title character varying(100)
public
datasetscleaned
id integer
creatoruserid integer
forumid integer
creationdate timestamp without time zone
lastactivitydate timestamp without time zone
totalviews integer
totaldownloads integer
totalvotes integer
totalkernels smallint
public
competitiontags
id integer
competitionid integer
tagid integer
public
users
id integer
username character varying(60)
displayname character varying(260)
registerdate date
performancetier smallint
country character varying(40)
public
tags
id integer
parenttagid double precision
name character varying(50)
slug character varying(85)
fullpath character varying(95)
description character varying(300)
datasetcount integer
competitioncount integer
kernelcount integer
public
datasettags
id integer
datasetid integer
tagid integer
public
userfollowers
id integer
userid integer
followinguserid integer
creationdate date
public
userachievements
id integer
userid integer
achievementtype character varying(15)
tier smallint
tierachievementdate character varying(30)
points integer
currentranking double precision
highestranking double precision
totalgold smallint
totalsilver smallint
totalbronze smallint
public
organizations
id smallint
name character varying(60)
slug character varying(60)
creationdate date
description text
public
userorganizations
id smallint
userid integer
organizationid smallint
joindate date
public
submissionscleaned
id integer
submitteduserid double precision
teamid integer
submissiondate date
isafterdeadline boolean
publicscoreleaderboarddisplay double precision
privatescoreleaderboarddisplay double precision