DATABASE COURSE

SQL & Data Modeling Sprint – 2-Day Capstone Project

Edited by :**Hoor Sultan Rashid Alkaabi**

Supervisor : **Fatma Almamari**

NOTE : SQL FILE AT GITHUB

Table of Contents :

Contents

• Research

Introduction :

Why Are SQL and Data Modeling Essential Competencies in AI and Data Science:

Data serves as the basis for training models and making judgments in the rapidly developing domains of artificial intelligence (AI) and data science However, how data is organized, saved, and retrieved has a significant impact on its utility. SQL (Structured Query Language) and data modeling become crucial tools at this point. Professionals with these abilities can arrange, retrieve, and handle data in ways that guarantee effectiveness, precision, and scalability—elements essential to the success of AI systems.

How Do Data Retrieval and Storage Impact AI/ML Training Outcomes:

The speed and accuracy of training machine learning (ML) models are directly impacted by effective data retrieval and storage. Data that is poorly organized can result in: training pipelines that are slow because of duplicated data or ineffective joins. longer preprocessing time, which causes delays and raises computation expenses. Real-world For instance, Uber's ML platform Michelangelo trains models for ETAs and driver demand predictions by extracting and joining structured data from several sources using SQL. Effective data modeling improves model performance by lowering latency in retrieving past trip data . (Uber Engineering, 2017)

Technical Debt Is Reduced by Clean, Well-Modeled Data :

The future expenses incurred by implementing short-term, temporary fixes in systems rather than long-term solutions are referred to as technical debt. Technical debt in AI/ML pipelines frequently originates from:

unclear schemas or unrecorded data connections, model inputs that are broken by inconsistent data formats .

You can build scalable, reusable data pipelines by developing a solid data model. When SQL is used with well-defined schemas, the data is guaranteed to be consistent, verified, and auditable.

Actual world For instance, Google's ML Engineering Guide ("Rules of Machine Learning") cautions that unstable data pipelines, not model programming, are the primary cause of most production model failures. Teams lower the chance of errors and rework by utilizing standardized schemas and SQL queries with validation.

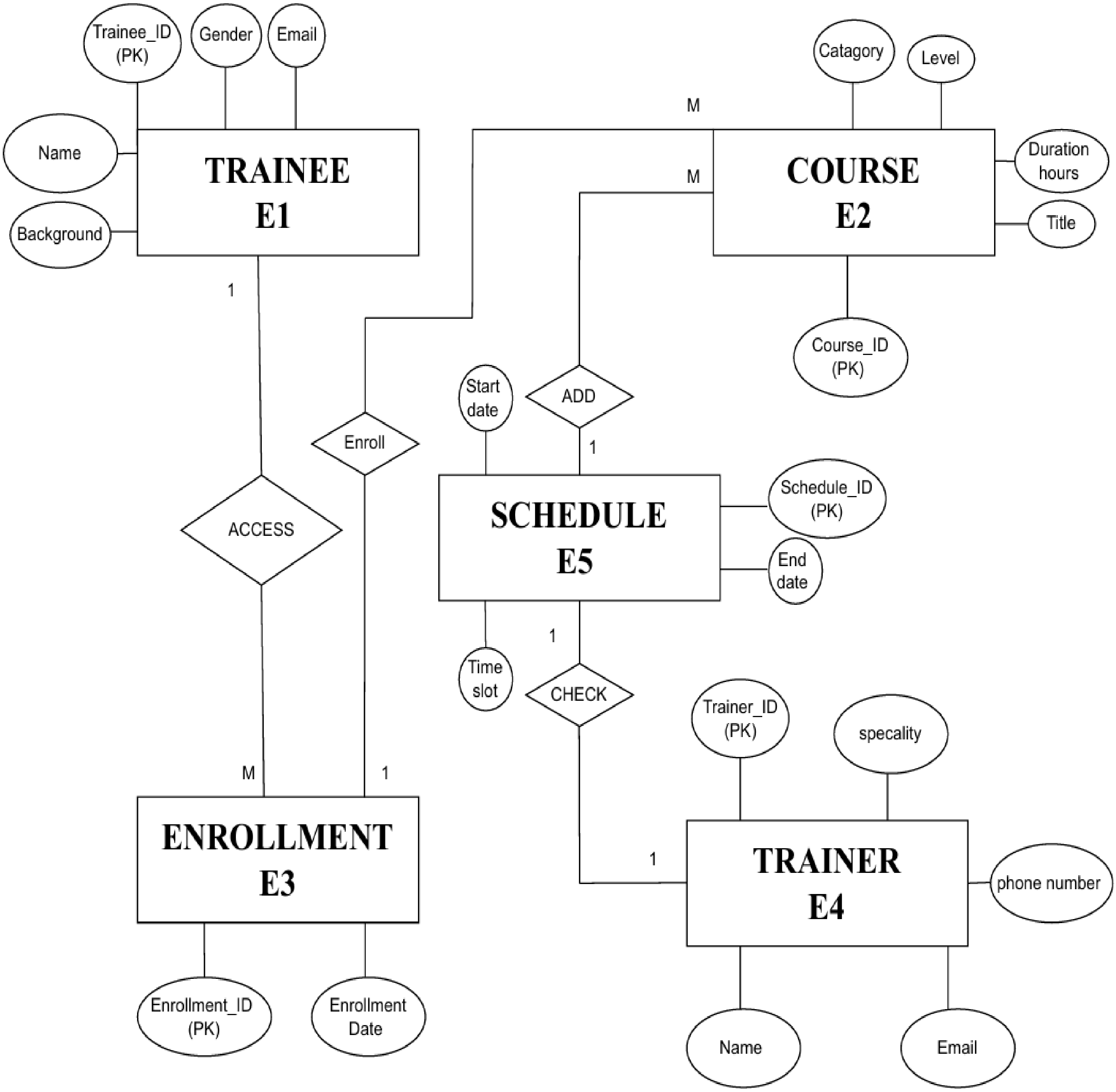
Examples of Using Structured Databases for Governance, Monitoring, and Auditing :

Data governance is crucial for AI systems, particularly those employed in government, healthcare, and finance . This comprises: who can see or change the data is known as access control ,Audit trails: When and by whom was the data altered or queried, Data lineage: The origins and processing methods of the data , SQL databases come with logging, authorization, and query auditing built right in. Actual world For instance, SQL dashboards are used by Airbnb to track the movement of data via machine learning pipelines. To keep track of any modifications, they strictly enforce version control on table schemas.

Thinking Back: Why This Is Important to My Future in AI: This course changed how I view data—I've learned that good models rely on clean, well-structured data. Whether building a recommendation engine or an enrollment system, defining clear relationships and retrieving data efficiently

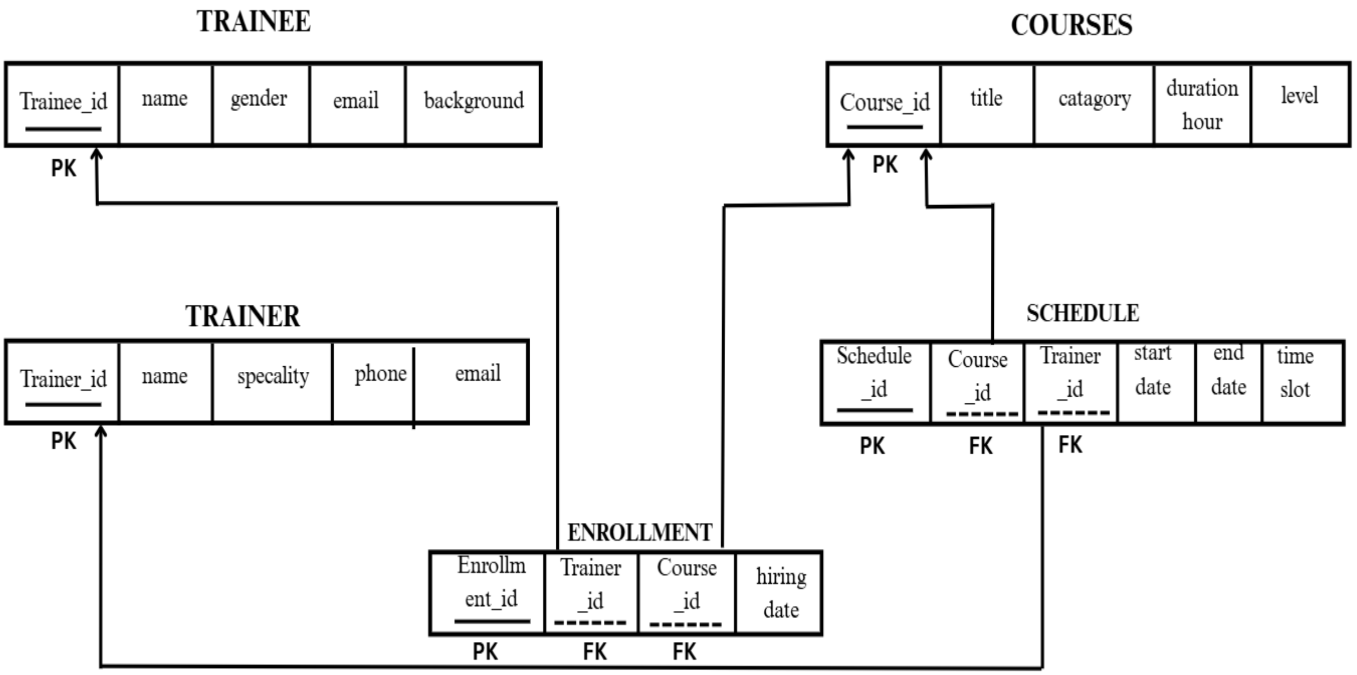
- Design

ERD Diagram



Picture 1: (ERD) Done with drawio browser

- ER – Mapping :



Picture 2: (Mapping) Map Done with drawio browser

Primary Key (PK)

- A Primary Key uniquely identifies each record (row) in a table.
- It must contain unique values and cannot contain NULLs.
- Each table can have only one Primary Key.
- Often used as the main reference for linking with other tables.

Foreign Key (FK)

- A Foreign Key is a field (or set of fields) in one table that refers to the Primary Key in another table.
- It is used to establish and enforce a link between the data in two tables.
- Can contain duplicate values and NULLs (unless restricted).
- Helps maintain referential integrity between related tables.
- A table can have multiple foreign keys.

• Implement AT SQL

DDL (Data Definition Language) :

Command	**Purpose**
CREATE :	Creates a new table, database, index, view, etc.
ALTER :	Changes the structure of an existing object (e.g., add a column).
DROP :	Deletes an object from the database.
TRUNCATE:	Removes all records from a table but keeps its structure.
RENAME:	Changes the name of an object.

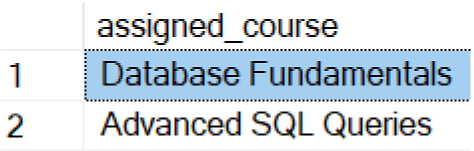
DML (Metadata Lock in MySQL) :

Purpose of MDL:

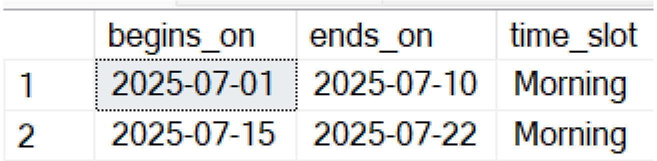
- Ensures data consistency.
- Prevents conflicts between reading and modifying a table.

- Query

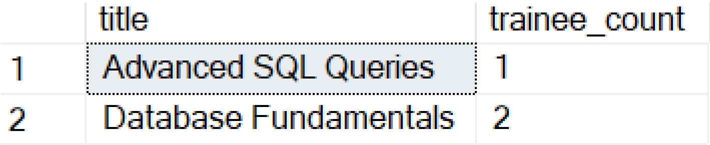
Trainer Perspective



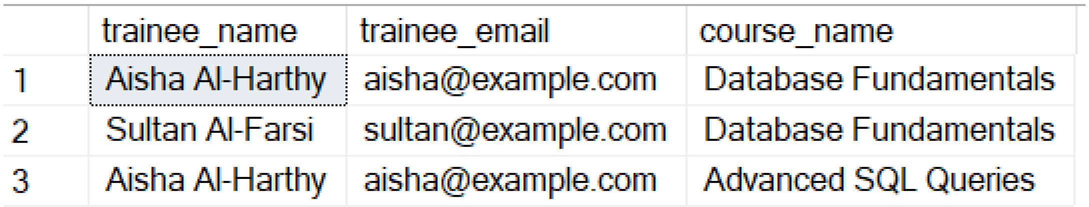
	assigned_course
1	Database Fundamentals
2	Advanced SQL Queries



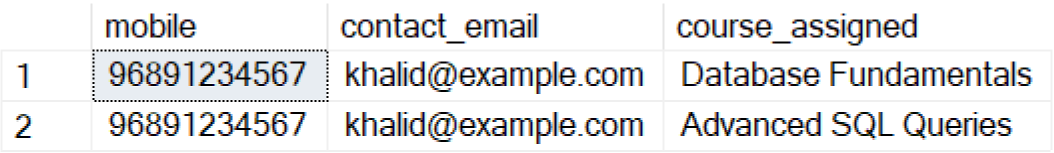
	begins_on	ends_on	time_slot
1	2025-07-01	2025-07-10	Morning
2	2025-07-15	2025-07-22	Morning



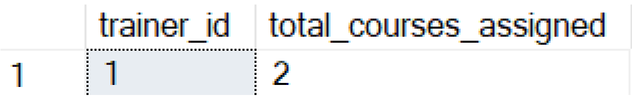
	title	trainee_count
1	Advanced SQL Queries	1
2	Database Fundamentals	2



	trainee_name	trainee_email	course_name
1	Aisha Al-Harthy	aisha@example.com	Database Fundamentals
2	Sultan Al-Farsi	sultan@example.com	Database Fundamentals
3	Aisha Al-Harthy	aisha@example.com	Advanced SQL Queries

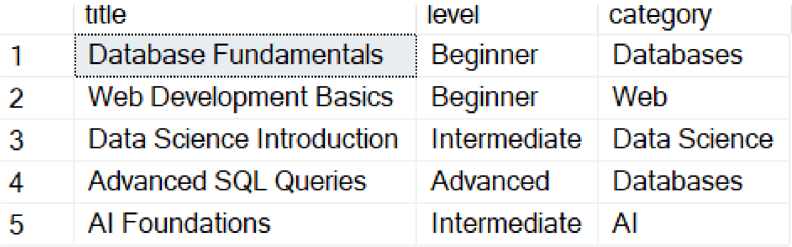


	mobile	contact_email	course_assigned
1	96891234567	khalid@example.com	Database Fundamentals
2	96891234567	khalid@example.com	Advanced SQL Queries

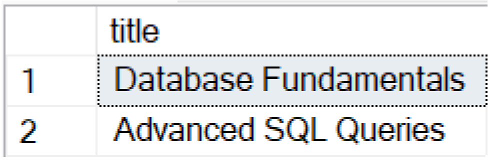


	trainer_id	total_courses_assigned
1	1	2

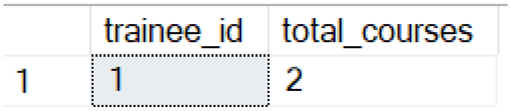
Trainee Perspective



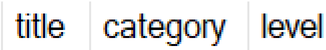
	title	level	category
1	Database Fundamentals	Beginner	Databases
2	Web Development Basics	Beginner	Web
3	Data Science Introduction	Intermediate	Data Science
4	Advanced SQL Queries	Advanced	Databases
5	AI Foundations	Intermediate	AI



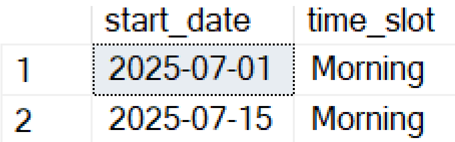
	title
1	Database Fundamentals
2	Advanced SQL Queries



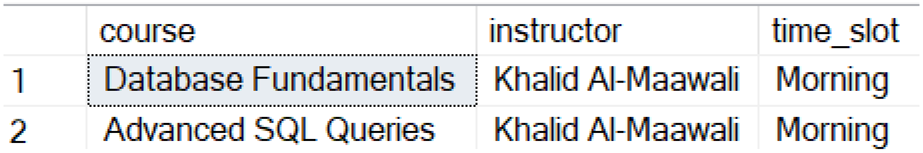
	trainee_id	total_courses
1	1	2



title	category	level

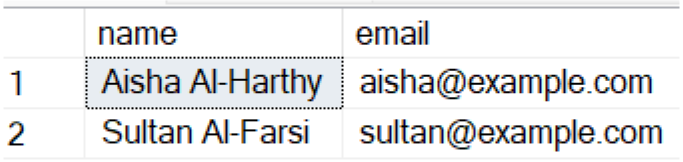


	start_date	time_slot
1	2025-07-01	Morning
2	2025-07-15	Morning

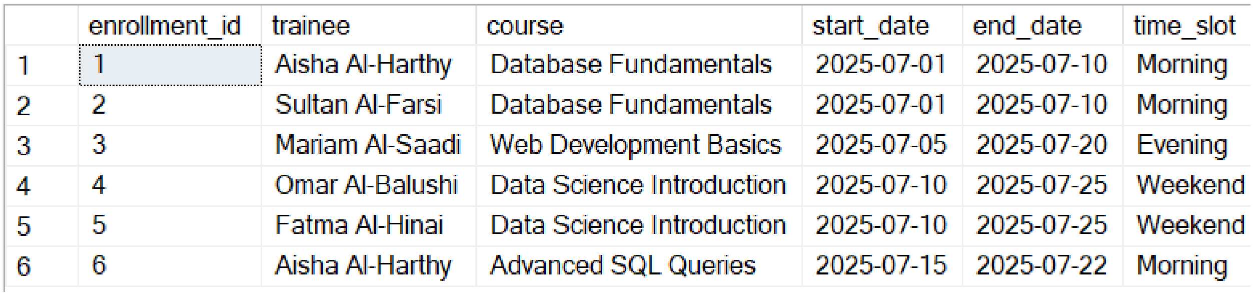


	course	instructor	time_slot
1	Database Fundamentals	Khalid Al-Maawali	Morning
2	Advanced SQL Queries	Khalid Al-Maawali	Morning

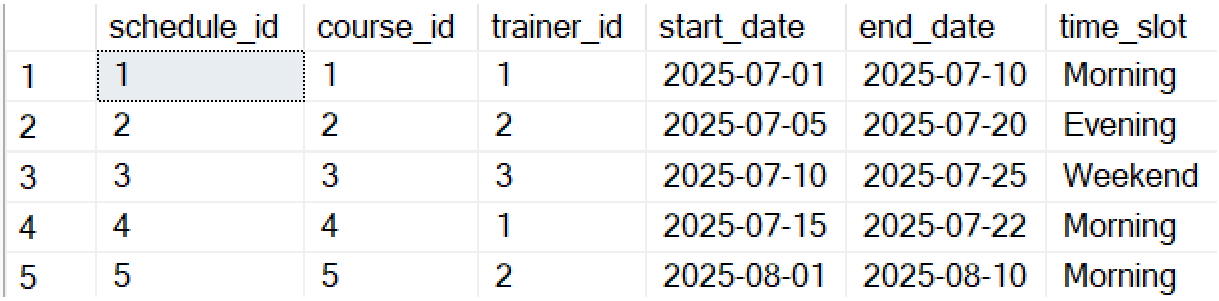
Admin Perspective



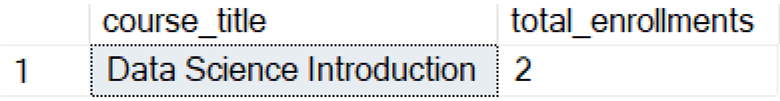
	name	email
1	Aisha Al-Harthy	aisha@example.com
2	Sultan Al-Farsi	sultan@example.com



	enrollment_id	trainee	course	start_date	end_date	time_slot
1	1	Aisha Al-Harthy	Database Fundamentals	2025-07-01	2025-07-10	Morning
2	2	Sultan Al-Farsi	Database Fundamentals	2025-07-01	2025-07-10	Morning
3	3	Mariam Al-Saadi	Web Development Basics	2025-07-05	2025-07-20	Evening
4	4	Omar Al-Balushi	Data Science Introduction	2025-07-10	2025-07-25	Weekend
5	5	Fatma Al-Hinai	Data Science Introduction	2025-07-10	2025-07-25	Weekend
6	6	Aisha Al-Harthy	Advanced SQL Queries	2025-07-15	2025-07-22	Morning



	schedule_id	course_id	trainer_id	start_date	end_date	time_slot
1	1	1	1	2025-07-01	2025-07-10	Morning
2	2	2	2	2025-07-05	2025-07-20	Evening
3	3	3	3	2025-07-10	2025-07-25	Weekend
4	4	4	1	2025-07-15	2025-07-22	Morning
5	5	5	2	2025-08-01	2025-08-10	Morning



	course_title	total_enrollments
1	Data Science Introduction	2

Issues Faced :

ISSUE	SOLUTION
Not aggregated with COUNT(), MAX() SELECT trainee_id, course_id, -- ERROR: course_id is not aggregated or in GROUP BY COUNT(course_id) AS total_courses FROM Enrollment GROUP BY trainee_id HAVING trainee_id = 1;	Either add course_id to the GROUP BY (which changes the meaning of the query), or Remove course_id from SELECT SELECT trainee_id, COUNT(course_id) AS total_courses FROM Enrollment GROUP BY trainee_id HAVING trainee_id = 1;
ERROR: column reference "name" is ambiguous Incorrect Join Conditions No Enrollments for Trainer’s Courses	Verify table schemas and use proper join keys that represent the relationships. Use LEFT JOIN Enrollment and LEFT JOIN Trainee to include courses with zero trainees. LEFT JOIN Enrollment en ON en.course_id = cr.course_id LEFT JOIN Trainee trn ON trn.trainee_id = en.trainee_id

Issues Table 1:

Issue	**Fix**
course_id missing	Add course_id in the insert OR make it auto-increment
Possible duplication	Choose a unique ID or check existing rows
Mismatch in column count	Add course_id in the column list

References :

Uber Engineering. (2017, September 5). *Michelangelo: Uber's machine learning platform*. Uber Engineering Blog. https://eng.uber.com/michelangelo-machine-learning-platform/ (Uber Engineering, 2017)

Google Developers. (n.d.). *Rules of machine learning: Best practices for ML engineering*. https://developers.google.com/machine-learning/guides/rules-of-ml

Airbnb Engineering & Data Science. (2018, August 15). *Data quality at Airbnb*. Medium. https://medium.com/airbnb-engineering/data-quality-at-airbnb-d8b7e1d6e4b3

Forrester, J. (2021, May 12). *Why data scientists spend 80% of their time cleaning data*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2021/05/12/why-do-data-scientists-spend-80-of-their-time-cleaning-data/

Superconductive. (n.d.). *Managing data quality with expectations*. Great Expectations Documentation. https://docs.greatexpectations.io/

Learning Outcome :

Basics

- Understand tables, rows, columns, keys
- Know relational database concepts

DDL (Definition)

- CREATE, ALTER, DROP tables
- Set PRIMARY KEY, FOREIGN KEY

DML (Manipulation)

- INSERT – add
- UPDATE – edit
- DELETE – remove

DQL (Query)

- SELECT, WHERE, ORDER BY, DISTINCT

Aggregate Functions

- COUNT, SUM, AVG, MAX, MIN
- Use with GROUP BY, HAVING

Joins

- INNER, LEFT, RIGHT, FULL JOIN

Subqueries

- Nested queries inside SELECT, WHERE, etc.

Constraints

- NOT NULL, UNIQUE, DEFAULT, CHECK

Practice

- Real-world database scenarios
- Write and optimize queries