26-06-2025

The dataset **"Rain Prediction Train.csv"** contains 100,000 rows and 23 columns. It includes weather data such as temperature, humidity, pressure, wind conditions, and labels like `RainToday` and `RainTomorrow`.

Resultant files after applying Normalization:
https://drive.google.com/drive/folders/1nLbyoS8sGbv2AZMCr5RSfgu0PiEmw3J1?usp=sharing

Normalization 1NF → 3NF

We'll now walk through **normalizing this dataset** by breaking it into multiple well-structured tables to eliminate redundancy and ensure data integrity.

Unnormalized Data Example

From the dataset, consider the following sample attributes:

- `Date`, `Location`, `MinTemp`, `MaxTemp`, `Rainfall`, `RainToday`, `RainTomorrow`
- `WindGustDir`, `WindGustSpeed`, `WindDir9am`, `WindSpeed9am`, `WindDir3pm`, `WindSpeed3pm`

There are **repeating groups and mixed facts**, which we must separate through normalization.

1NF – First Normal Form

Goal: Eliminate repeating groups and ensure each field contains atomic values.

Issues Fixed:

- No column should have multiple values in one cell (not seen here).
- Ensure uniqueness with a primary key — we can assume `(Date, Location)`.

Sample 1NF Table:

Date	Location	MinTemp	MaxTemp	Rainfall	WindGustDir	WindGustSpeed	RainToday
07-02-2014	CoffsHarbour	17.7	25.9	2.2	NNE	31.0	Yes

2NF – Second Normal Form

Goal: Remove **partial dependencies** (non-key columns depending on part of a composite key).

Assumption: Composite Key = (Date, Location)

Issues Fixed:

- Attributes like `MinTemp`, `MaxTemp`, `Rainfall` depend fully on the composite key — OK.
- But some attributes (like `Location` details, city-specific climate averages if available) might be independent of Date → Move to separate tables.

Split Into:

Table 1: WeatherObservations

ObservationID	Date	Location	RainToday	RainTomorrow
1	07-02-2014	CoffsHarbour	Yes	No

Table 2: WeatherMetrics

ObservationID	MinTemp	MaxTemp	Rainfall	WindGustDir	WindGustSpeed
1	17.7	25.9	2.2	NNE	31.0

3NF – Third Normal Form

Goal: Remove **transitive dependencies** (non-key attribute depending on another non-key attribute).

Issues Fixed:

- If Location has fixed attributes like Region or Elevation (not in this dataset), they must be extracted.

Add Separate Location Table (if extra metadata is available):

Table 3: Locations

LocationID	Location
L1	CoffsHarbour

Table 4: WeatherObservations Updated

ObservationID	Date	LocationID	RainToday	RainTomorrow
1	07-02-2014	L1	Yes	No

Summary of Tables After 3NF

1. **WeatherObservations** — ObservationID, Date, LocationID, RainToday, RainTomorrow
2. **WeatherMetrics** — ObservationID, MinTemp, MaxTemp, Rainfall, WindGustDir, WindGustSpeed, etc.
3. **Locations** — LocationID, LocationName

ACID Properties

Atomicity – Entire transaction succeeds or fails completely

Consistency – Ensures data remains valid after transactions

Isolation – Transactions occur independently

Durability – Committed changes are permanent

SQL Isolation Levels

Isolation Level	Dirty Read	Non-Repeatable Read	Phantom Read
Read Uncommitted	Yes	Yes	Yes
Read Committed	No	Yes	Yes
Repeatable Read	No	No	Yes
Serializable	No	No	No

NoSQL vs Relational Databases

Feature	MongoDB	Cassandra	RDBMS
Type	Document database	Wide-column database	Relational database
Schema	Schema-less	Semi-structured	Fixed schema
Consistency	Tunable	Tunable	Strong
Availability	High	Very High	Medium to High
Scalability	Horizontal	Horizontal	Mostly vertical
Transactions	Limited support	Lightweight	Full support

Advanced Normalization – BCNF and 4NF

Boyce-Codd Normal Form (BCNF)

A table is in BCNF if every determinant is a candidate key

Example:

If Department determines Location but Department is not a candidate key, then not in BCNF

EmpID	Department	Location
E01	HR	Floor1

Fix: Separate department and location into another table

Fourth Normal Form (4NF)

Remove multi-valued dependencies

Example:

Student	Course	Hobby
John	Math	Chess
John	Math	Football

Fix: Split into two tables

- StudentCourse: Student, Course
- StudentHobby: Student, Hobby

Data Modeling Tools

Tool	Description
Lucidchart	Web-based tool for ER diagrams and schemas
DBML	Code-based schema modeling tool
Draw.io	Free diagram tool, supports ER diagrams
Vertabelo	Online database modeler with reverse engineering
MySQL Workbench	Visual schema designer and SQL generator