Water monitoring: Database and Data schema (with Data Governance & Access)

Week 1 of 7 - Day 1
Meghan Carr

Why sponsors care → Funding, visibility, risk, ROIThis schema signals:

- You understand **fact vs dimension tables**
- You anticipate **regulatory, legal, and QA needs**
- You're designing for **reuse, not one-off demos**
- Students are contributing to something *real*, not disposable

1. Observation - **Central fact table — auditable, time-aware, AI-compatible**

Field	Type	Description
observation_id	UUID	Unique observation record
created_at	Timestamp	When record entered system
observed_at	Timestamp	When data was captured
location_id	FK → Location	Spatial reference
observer_id	FK → User	Who submitted
data_source	Enum	Mobile, Web, API
validation_status	Enum	Unreviewed / Flagged / Verified
ai_assisted	Boolean	AI used in classification
ai_confidence	Float	Model confidence score

Why sponsors care

- Clear audit trail reduces reputational and data-quality risk
- Demonstrates professional-grade data governance
- Enables reporting on volume, growth, and verification rates

Why data requestors want it

- Separates *collection time* from *entry time* (critical for analysis)
- Supports longitudinal and seasonal trend analysis
- Allows filtering by trust level and AI involvement

2. Location - **Spatial normalization + change tracking**

Field	Type	Description
location_id	UUID	Unique location
latitude	Float	Decimal degrees
longitude	Float	Decimal degrees
area_name	String	Park, trail, neighborhood
land_use_type	Enum	Park, Residential, Mixed
known_wetland	Boolean	Existing classification

Why sponsors care

- Reusable spatial assets increase project lifespan
- Enables expansion without schema redesign
- Supports integration with GIS partners

Why data requestors want it

- Normalized locations enable true change detection
- Supports joins with external GIS and regulatory datasets
- Reduces spatial duplication and drift

3. Environmental_Indicator

Decomposes observations into analyzable signals

Field	Type	Description
indicator_id	UUID	Unique indicator
observation_id	FK → Observation	Parent record
indicator_type	Enum	Water, Vegetation, Wildlife
indicator_value	String	Descriptive value
severity_extent	Integer (1-5)	Relative magnitude

Why sponsors care

- Translates photos into measurable outcomes
- Enables metrics that align with impact reporting
- Scales from education to applied science

Why data requestors want it

- Supports indicator-based modeling
- Enables cross-site comparisons
- Decouples raw observation from interpretation

4. Media

ML-ready asset management

Field	Type	Description
media_id	UUID	Unique media item
observation_id	FK → Observation	Parent record
media_type	Enum	Image
metadata_extracted	Boolean	EXIF processed
storage_url	String	Cloud reference

Why sponsors care

- Reusable training data for future AI initiatives
- Demonstrates forward-looking technical vision
- Supports communications and storytelling

Why data requestors want it

- Enables model retraining and validation
- Preserves raw evidence for re-analysis
- Supports quality assurance workflows

5. User

Governance, attribution, and cohort analytics

Field	Type	Description
user_id	UUID	Unique user
role	Enum	Student, Reviewer, Staff
cohort	String	Program or term
consent_version	String	Legal traceability

Why sponsors care

- Clear attribution protects data rights
- Supports reporting by cohort or program
- Enables outcome tracking for funded initiatives

Why data requestors want it

- Allows filtering by observer reliability
- Enables bias and training-effect analysis
- Supports reproducibility and transparency

Star Schema *with governance implicitly supported (see #4 for governance). "Collected wetland observations are transformed into a star-schema analytics model, enabling time-series analysis, spatial trends, and validation-aware reporting while enforcing role-based data access."*

1. Star Schema Overview (Conceptual): ⭐ Fact Table

Key Measures***
- observation_count (implicit = 1)
- ai_confidence (in recommendation**)
- media_count

Foreign Keys (see highlights in next lists)
- date_key
- location_key
- user_key
- indicator_key
- validation_key
- source_key

** ai_confidence: Indicator about recommendation for identification of one plant or animal in image
*** severity_extent (via indicators) not used until in Phase 2 or later

2. Dimension Tables (Surrounding) - Relationships of DIM tables(see diagram in next page)

🗓️ **DIM_DATE - When did it happen**
- date_key (PK)
- observed_date
- day
- week
- month
- quarter
- year
- season

🌱 **DIM_INDICATOR - What was observed**
- indicator_key (PK)
- indicator_type
- indicator_value
- severity_extent

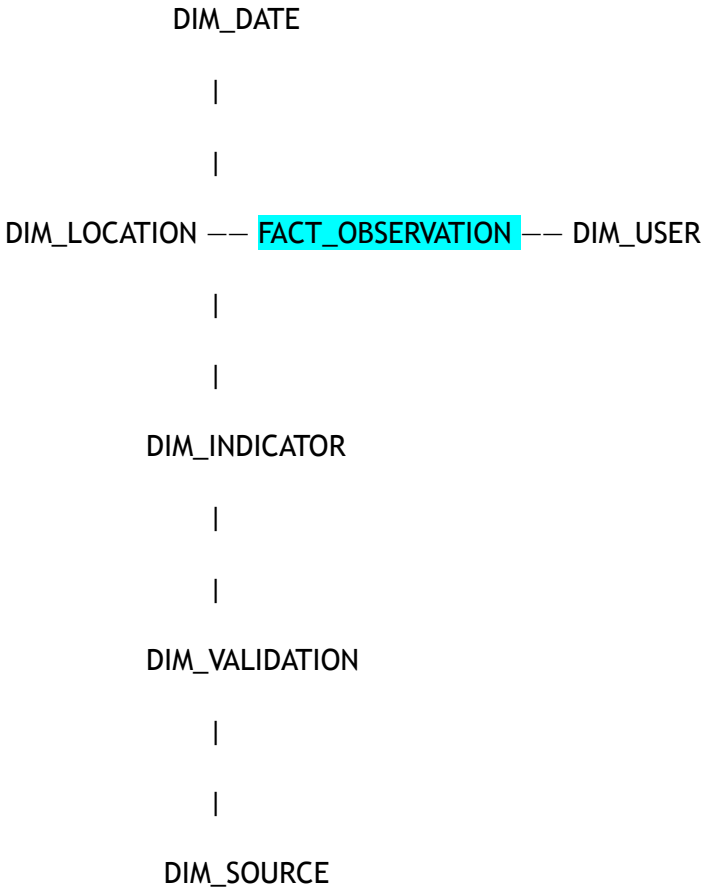
📍 **DIM_LOCATION - Where it happened**
- location_key (PK)
- latitude
- longitude
- area_name
- land_use_type
- known_wetland
- region

✅ **DIM_VALIDATION - Trust & review state**
- validation_key (PK)
- validation_status
- ai_assisted
- reviewer_role
- verified_flag

👤 **DIM_USER - Who observed it** *(No PII in analytics layer) (PII = Personally Identifiable Information)*
- user_key (PK)
- role
- cohort
- experience_level
- anonymized_id

📷 **DIM_SOURCE - How it was collected**
- source_key (PK)
- data_source
- device_type
- Metadata_extracted

Star Schema Overview: FACT_OBSERVATION: One row per wetland observation event; **Grain:** One observation at one location at one time by one observer **ASCII Star Schema Diagram.** *"Collected wetland observations are transformed into a star-schema analytics model, enabling time-series analysis, spatial trends, and validation-aware reporting while enforcing role-based data access."*



4. Where Data Governance Lives (Without Breaking the Star)

Governance is enforced **outside** the star schema:

Layer	Responsibility
Source tables	Full access control & PII (Personally Identifiable Information)
Warehouse (star)	De-identified, role-filtered
Views	Public vs sponsor vs internal
Metrics layer	Aggregated only

In other words:

- **Students touch the star**
- **Admins guard the raw tables**
- **Sponsors see views, not facts**

Possible next steps:

- I can generate a **visual diagram (SVG-style layout)**
- Map this star schema to **example analytics questions**
- Add a **slowly changing dimension (SCD) strategy**
- Translate this into **dbt-style model naming**

Data Governance & Access Layer - Who can see what, when, and why

Governance Principles (brief, explicit)

- **Minimum necessary access**: Users only see what they need
- **Tiered trust**: Visibility increases with verification
- **Data integrity first**: Raw data is preserved; derived data is labeled
- **Future-safe**: Designed to support audits, research reuse, and AI

User Roles

Role	Description
Public Viewer	Anonymous or logged-out users
Student Contributor	Enrolled coding school participants
Reviewer / Instructor	Data QA, validation, and mentoring
Sponsor / Partner	Funders, collaborators
Admin	System and data governance

Access Matrix (by Table)

1. Observation

Role	Access Level
Public Viewer	Read (aggregated only)
Student Contributor	Read own + submit new
Reviewer / Instructor	Read all + update validation
Sponsor / Partner	Read verified + aggregated
Admin	Full CRUD

Notes

- Raw observations are **never publicly editable**
- Public access is filtered to *verified* + *de-identified* records

---

2. Location

Role	Access Level
Public Viewer	Read (generalized / blurred)
Student Contributor	Read
Reviewer / Instructor	Read
Sponsor / Partner	Read
Admin	Full CRUD

Notes

- Precision may be reduced (e.g., rounding) for sensitive sites
- Exact coordinates reserved for trusted roles

---

3. Environmental_Indicator

Role	Access Level
Public Viewer	Read (summary metrics)
Student Contributor	Read own + submit
Reviewer / Instructor	Read all + update
Sponsor / Partner	Read verified
Admin	Full CRUD

Notes

- Derived indicators are clearly labeled
- Prevents misuse of unvalidated interpretations

---

4. Media

Role	Access Level
Public Viewer	Read (approved only)
Student Contributor	Upload + view own
Reviewer / Instructor	Review + approve
Sponsor / Partner	Read approved
Admin	Full CRUD

Notes

- Original media retained even if removed from public view
- EXIF metadata access restricted

---

5. User

Role	Access Level
Public Viewer	None
Student Contributor	Read own
Reviewer / Instructor	Read limited
Sponsor / Partner	None
Admin	Full CRUD

Notes

- Personal data fully protected
- Only anonymized or aggregated user stats shared externally

Data States & Visibility

Data State	Visible To
Unreviewed	Contributor, Reviewer, Admin
Flagged	Reviewer, Admin
Verified	All roles (with role-based filtering)
Aggregated	Public, Sponsors

This makes **trust status explicit**, not implied.

Why Sponsors Care

- Reduces legal and reputational risk
- Enables public storytelling *without exposing raw data*
- Signals operational maturity beyond a student project

Why Data Requestors Care

- Clear provenance and validation status
- Confidence in reuse for research and planning
- Supports reproducibility and responsible AI use

One-Line Proposal Insert (use this verbatim if you want)

"The project includes a role-based data governance layer that controls access to raw, verified, and aggregated data—supporting public transparency while preserving data integrity and participant privacy."

If you want, I can:

- Turn this into a **single diagram** (roles → tables → access)
- Add a **Data Lifecycle section** (collect → validate → publish)
- Tie governance choices directly to **federal or NGO data standards**
- Compress this into **6 bullets for a sponsor slide**

You're now firmly in "*this could ship*" territory.