05_Data_Specs_and_Schemas

05 — Data Specs & Schemas

Station Schema (`stations`)

- `station_id` (text, pk) — NDBC code or synthetic id
- `name` (text)
- `lat` (float), `lon` (float)
- `provider` (enum: NDBC, ERDDAP, ERSST, ...)
- `first_obs` (timestamptz), `last_obs` (timestamptz)

Observation Schema (`buoy_obs`)

- `id` (bigserial, pk)
- `station_id` (fk stations.station_id)
- `time` (timestamptz, indexed)
- `sst_c` (float)
- `qc_flag` (int, 0=ok, >0 vendor-specific)
- `lat` (float), `lon` (float)
- `source` (text)

Indexes: `(station_id, time)`, `gist(lat, lon)` or btree on ranges.

JobRun (`job_runs`)

- `id`, `source`, `started`, `ended`, `status` (ok/failed), `rows_ingested`, `error`

Data Quality & QC

- Treat `MM` and sentinel values as nulls.
- Drop obviously bad SST (e.g., < -3°C or > 40°C) unless flagged for analysis.
- Track ingest counts vs vendor counts; alert on deltas > threshold.

Retention & Lineage

- Raw files cached 90 days; derived tiles rebuild on schedule.
- Provenance recorded in `job_runs` + artifact hashes.
- Metadata includes units, CRS (EPSG:3857 for tiles; WGS84 for points).

Units & Conversions

- Temperatures in °C; display toggle °C/°F in UI (Phase 2).