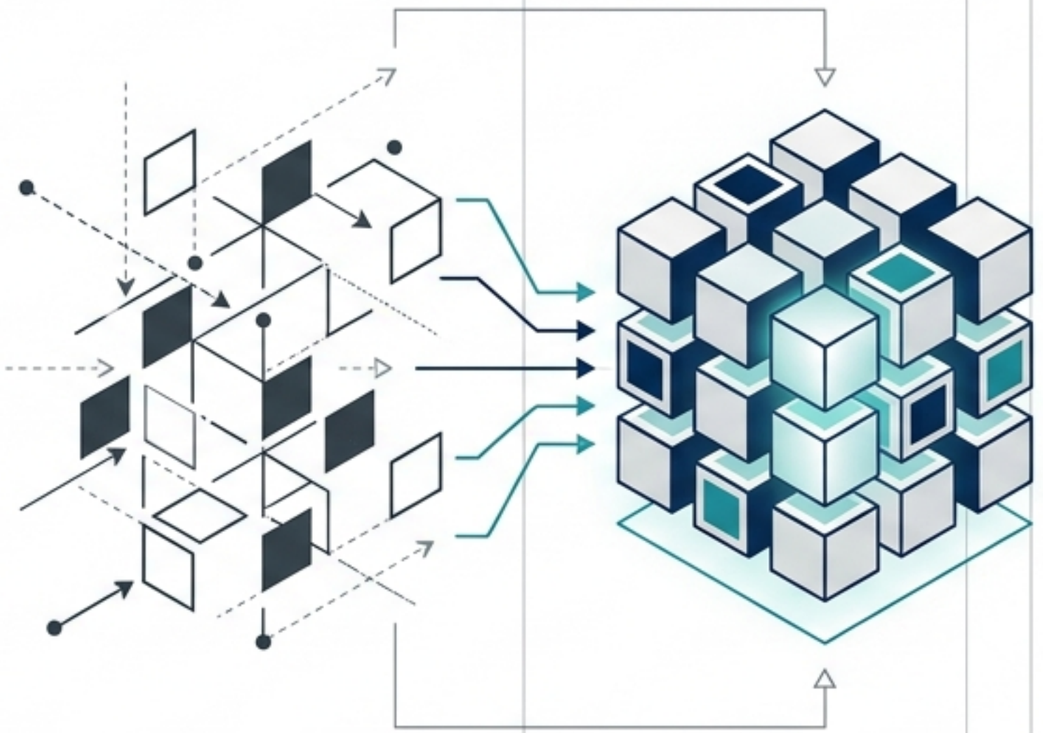
TECHNICAL BRIEFING // DATA ARCHITECTURE

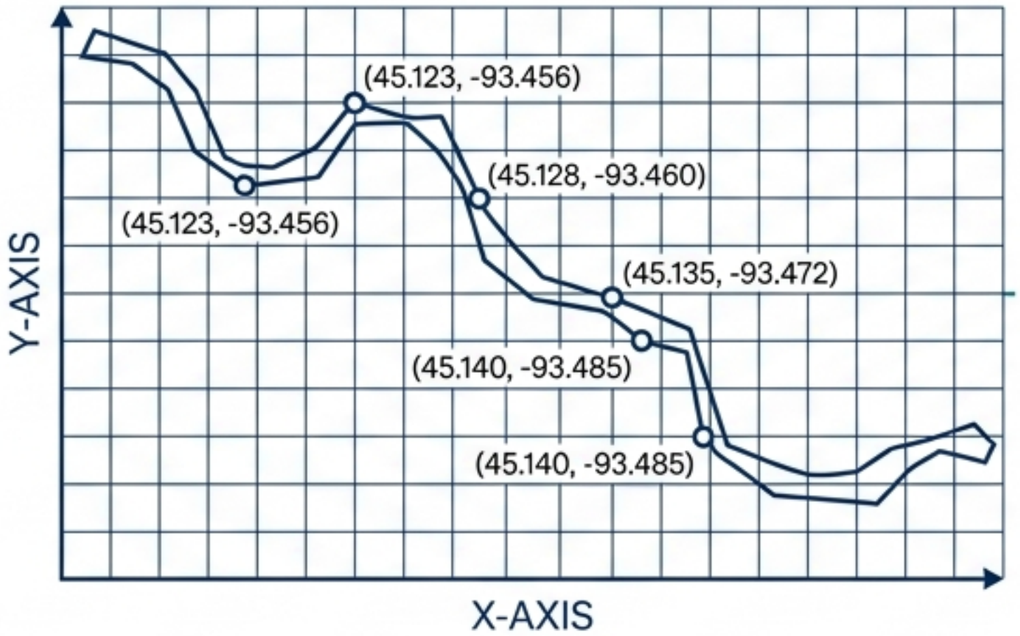
Modernizing Geospatial Data Formats for Hydrology

Transitioning government hydrological infrastructure from legacy local files to cloud-native, N-dimensional architectures.



THE ANATOMY OF A GEOSPATIAL FILE

COORDINATE GEOMETRIES



Spatial representation via explicit point, line, or polygon coordinates plotted on a defined grid system.

DESCRIPTIVE ATTRIBUTES

ATTRIBUTE	VALUE
RIVER NAME	Mississippi River
FLOW RATE	16,792 m^3/s
BASIN AREA	2,981,076 km^2

Contextual, non-spatial data linked to each geometric feature, providing essential information.

Geospatial files must synchronously store spatial coordinates alongside contextual data. Different formats are fundamentally optimized to balance spatial analysis, storage efficiency, and web sharing.

The Structural Limitations of the Legacy Shapefile (.shp)

Multi-File Dependency: Requires at least three mandatory files.

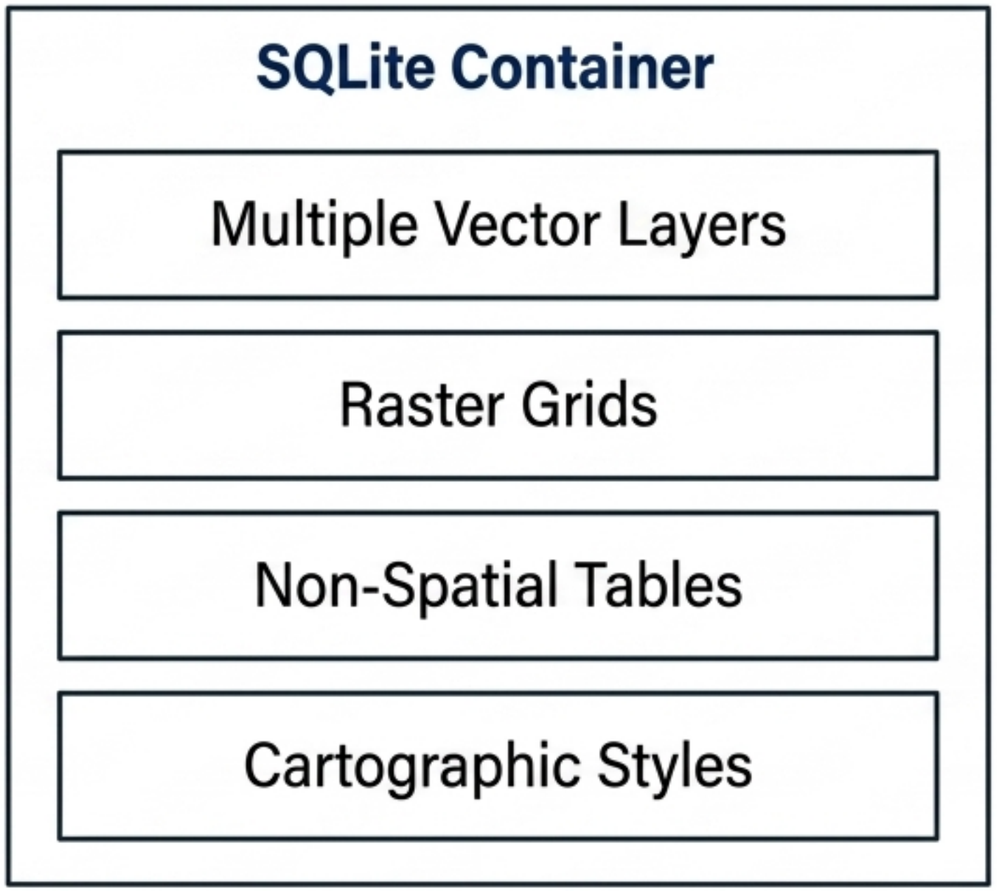
Truncated Metadata: Column headers are strictly limited to 10 characters (dBase IV legacy constraint).

Hard Size Caps: Total size of any individual component file cannot exceed 2 GB.

Inflexible Geometries: A single file cannot mix geometries (e.g., points and lines).

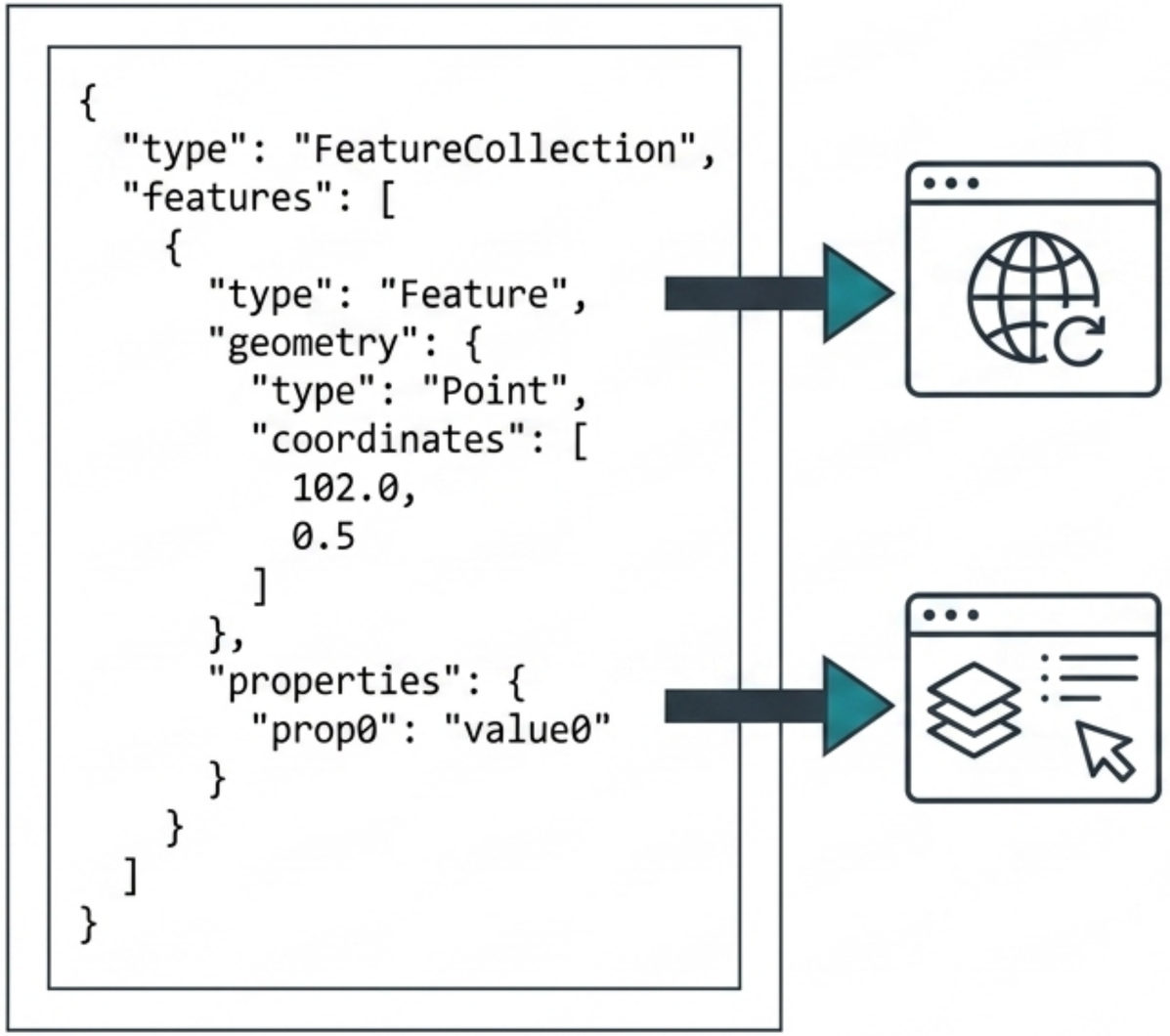
Poor Temporal Tracking: Lack of robust support for null values and coordinate time stamps, severely hindering time-series sensor data.

Consolidating Architecture with the Modern GeoPackage (.gpkg)



- **Definition:** The OGC-standardized modern replacement for legacy shapefiles.
- **Single-File Convenience:** Entire database stored in one easy-to-share file.
- **Infinite Capacity:** No arbitrary limits on column name lengths or table sizes.
- **Database Enabled:** Supports SQL triggers, views, and out-of-the-box spatial indexing for rapid large-scale querying.

Standardizing Web Exchange via GeoJSON



- **Definition:** An open, text-based format designed for web mapping applications and API data exchange.
- **Standardized CRS:** Hardcoded to always use WGS 84 (EPSG:4326) coordinates.
- **Web Native:** Parsed natively by browsers without intermediary processing.
- **The Constraint:** Because it represents geometries and attributes as plain text, file sizes balloon rapidly, making it highly inefficient for storing large datasets.

Tracking Simple Sensor Logs with Coordinate CSVs

ID	Timestamp	X (Longitude)	Y (Latitude)	Value
1	2023-10-26T10:00:00	-77.0365	38.8951	15.2
2	2023-10-26T10:05:00	-77.0370	38.8955	15.4
3	2023-10-26T10:10:00	-77.0375	38.8957	15.4
4	2023-10-26T10:15:00	-77.0400	38.8954	15.2
5	2023-10-26T10:20:00	-77.0425	38.8959	15.0
6	2023-10-26T10:25:00	-77.0420	38.8958	15.4
7	2023-10-26T10:30:00	-77.0450	38.8956	15.3

Delimited Text Provider

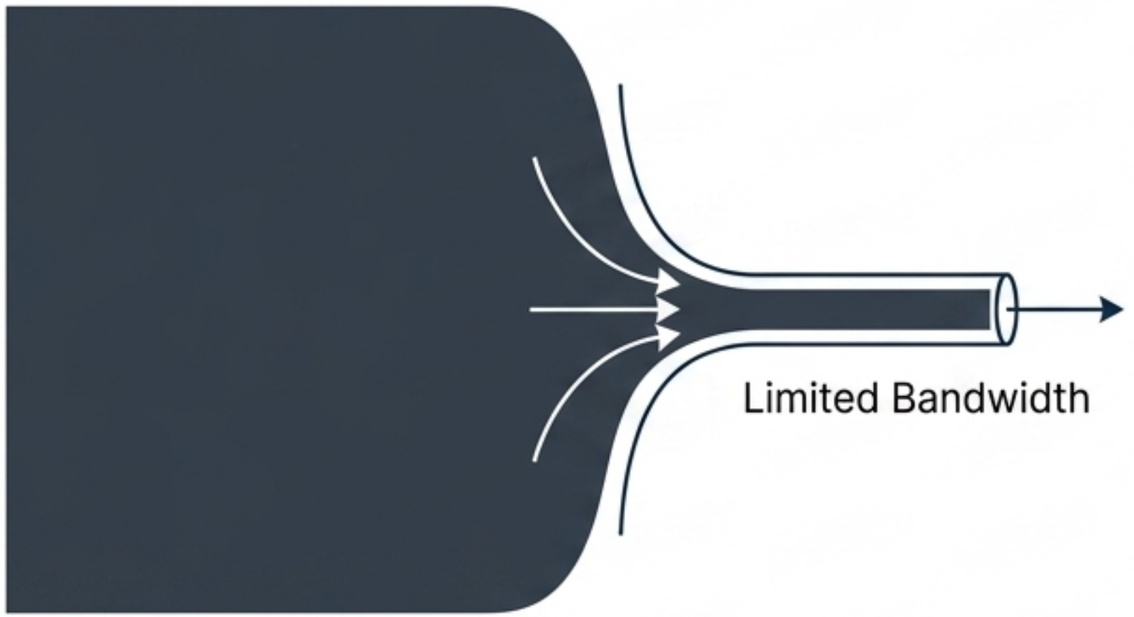
Definition:
Comma-Separated Values utilizing coordinate columns.

Hydrological Application:
The ubiquitous standard for exporting raw GPS logs, river gauge locations, and meteorological records.

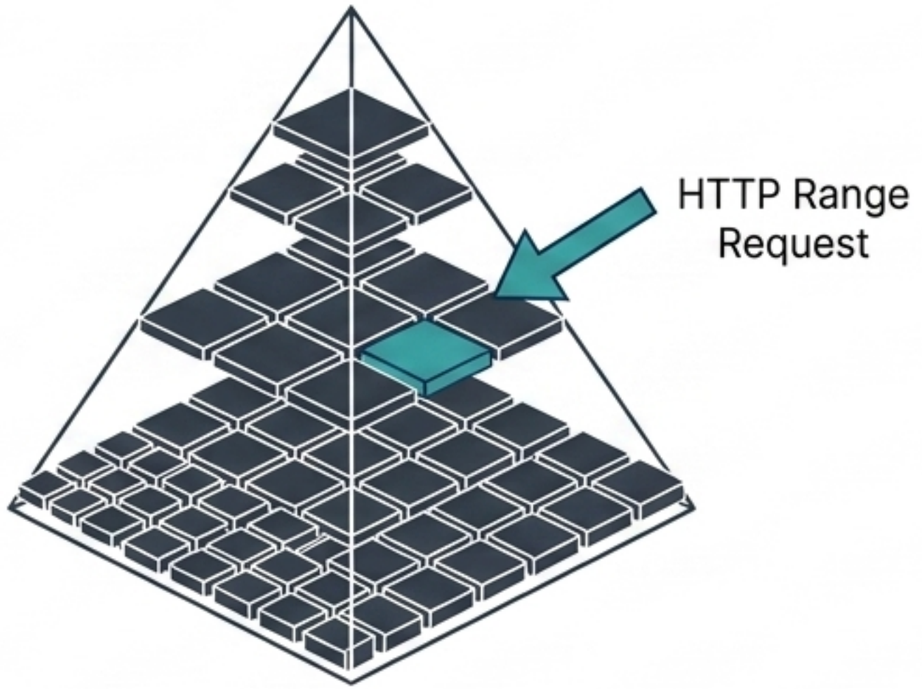
Integration:
Seamlessly parsed into vector point layers by standard GIS software.

Transforming Raster Access with Cloud-Optimized GeoTIFF (COG)

Legacy GeoTIFF



Cloud-Optimized GeoTIFF



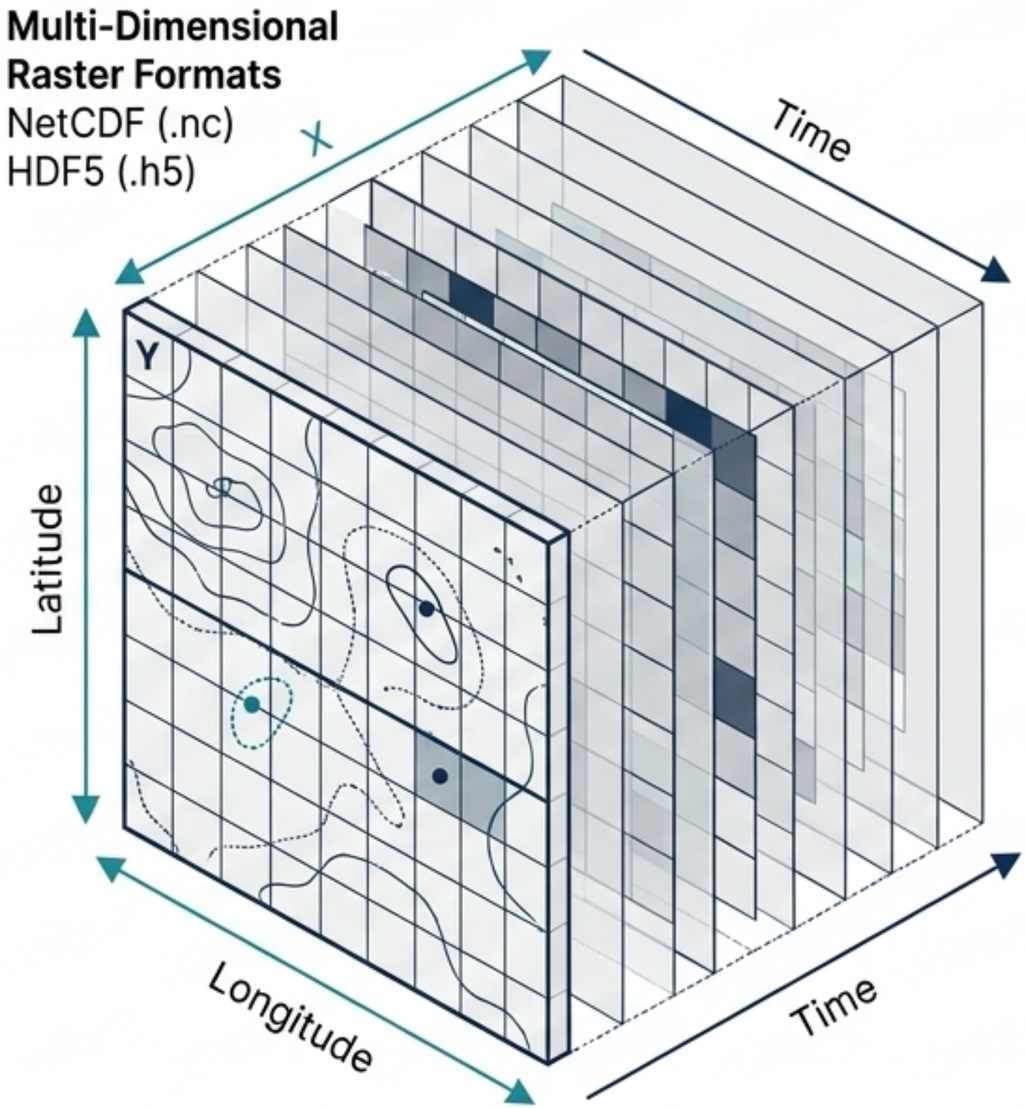
Definition: A standard GeoTIFF structurally reorganized for web environments.

Internal Tiling: Matrices are organized into localized tiles rather than continuous strips.

Overviews (Pyramids): Pre-rendered, lower-resolution versions are stored within the file.

HTTP Range Requests: Web clients download only the specific tile and resolution needed, eliminating multi-gigabyte downloads.

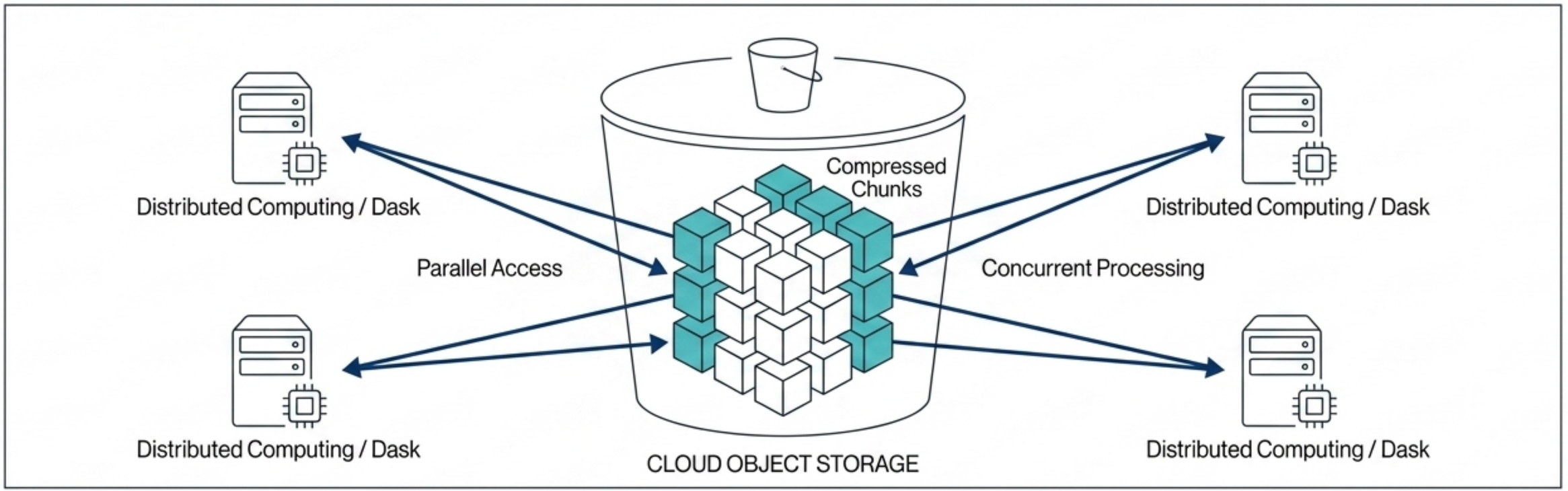
Structuring Time-Series Climate Arrays (NetCDF & HDF5)



NetCDF (.nc): Self-documenting, machine-independent format for scientific variables. Standard for multi-temporal climate/hydrology datasets (**CHIRPS** daily precipitation, **ERA5** reanalysis). Stores a decade of daily grids in one file.

HDF5 (.h5): Directory-like structure for massive, complex, heterogeneous datasets. Heavily utilized by space agencies for remote sensing missions (**GPM** precipitation, **SMAP** soil moisture, **MODIS** snow cover).

Enabling Distributed Cloud Computation via Zarr



Definition: The modern, cloud-native successor to NetCDF.

Structure: Stores chunked, compressed, N-dimensional arrays directly in cloud object storage.

Hydrological Application: Unlocks parallel access. **Enables distributed computing frameworks to process petabytes of climate and streamflow projections concurrently, rather than sequentially.**

Accelerating Massive Vector Queries (GeoParquet & FlatGeobuf)

Row-based Processing

Name	Geometry	Flow	Basin Area
'River X'	...	500 m^3/s	2000 km^2
'River Y'	...	500 m^3/s	1500 km^2
'River Z'	...	...	3000 km^2

Wastes compute scanning irrelevant fields.

Columnar Processing

Name	Geometry	Flow	Basin Area
'River X'	...	500 m^3/s	2000 km^2
'River Y'	...	500 m^3/s	1500 km^2
'River Z'	...	...	3000 km^2

Bypasses irrelevant data for 100x speed.

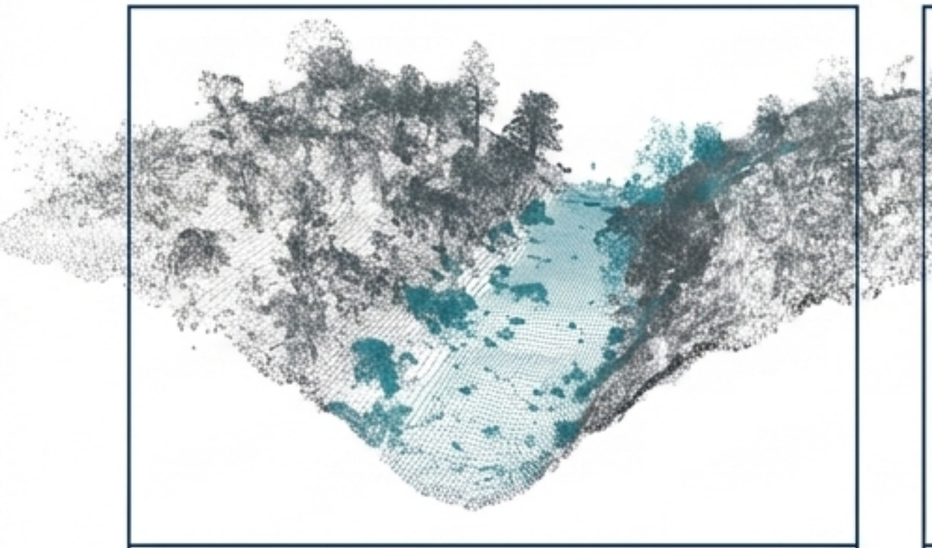
GeoParquet:

Columnar storage extension. Ideal for massive spatial vector tables (e.g., millions of global watershed boundaries). Achieves queries up to 100x faster than traditional databases by eliminating full-file scanning.

FlatGeobuf:

Binary vector format containing a packed R-Tree spatial index. Optimized for cloud streaming, fetching only river segments inside an active screen boundary without downloading the full layer.

Processing High-Resolution Terrain Point Clouds

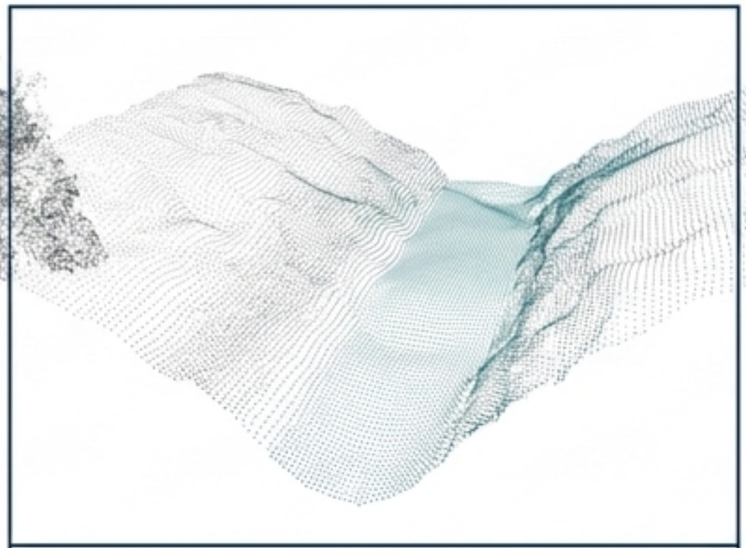


Stage 1: LAS (.las)

The foundational open binary format.

Records X, Y, Z coordinates alongside intensity, GPS time, and point classification (ground, water, vegetation).

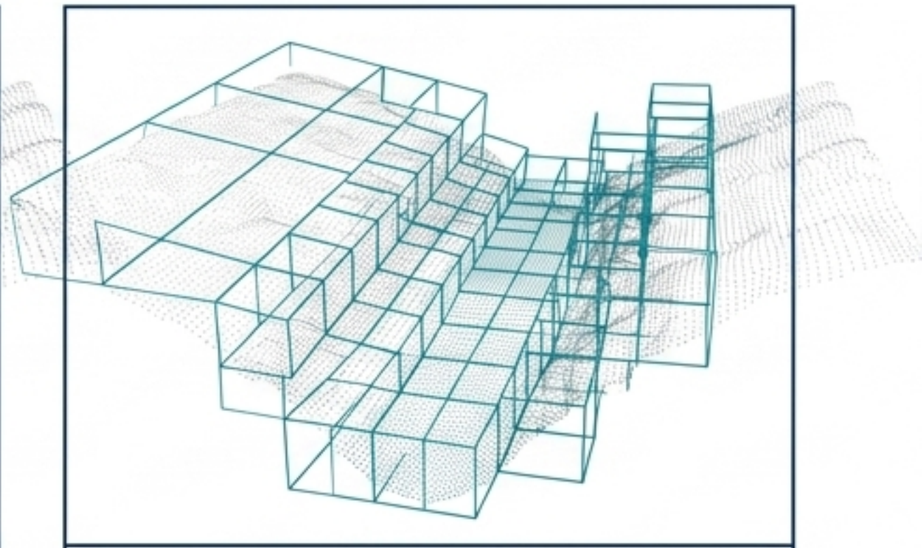
Essential for generating Bare-Earth Digital Terrain Models (DTMs).



Stage 2: LAZ (.laz)

The sharing standard.

Applies lossless compression to LAS files, reducing total file size by 70% to 90%.

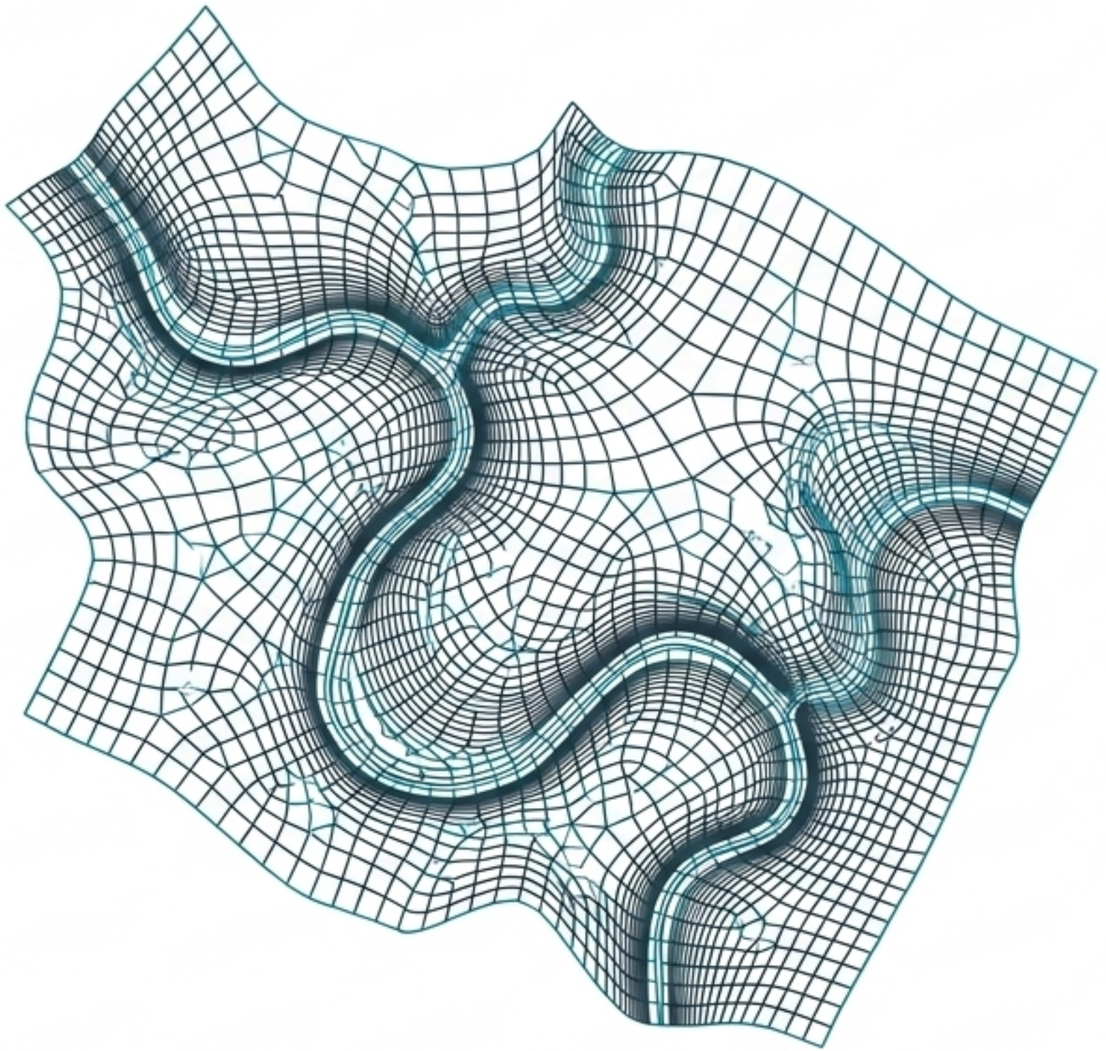


Stage 3: COPC (.copc.laz)

The cloud-streaming standard.

Reorganizes the LAZ file internally as a spatial octree, allowing dynamic streaming of river beds directly from cloud buckets based on the user's active view.

Simulating Flow Dynamics with 3D and Mesh Formats



HEC-RAS HDF5 Meshes

Structure: Embedded HDF5 files containing structured or unstructured computational mesh geometries.

Application: The engine for 2D hydraulic models **calculating** water surface elevations, flow velocities, and shear stress across complex terrain.

3D Tiles

Structure: OGC standard for streaming massive 3D geospatial content.

Application: Rendering complex structural models like dams, river gorges, or urban catch-systems interactively on web platforms.

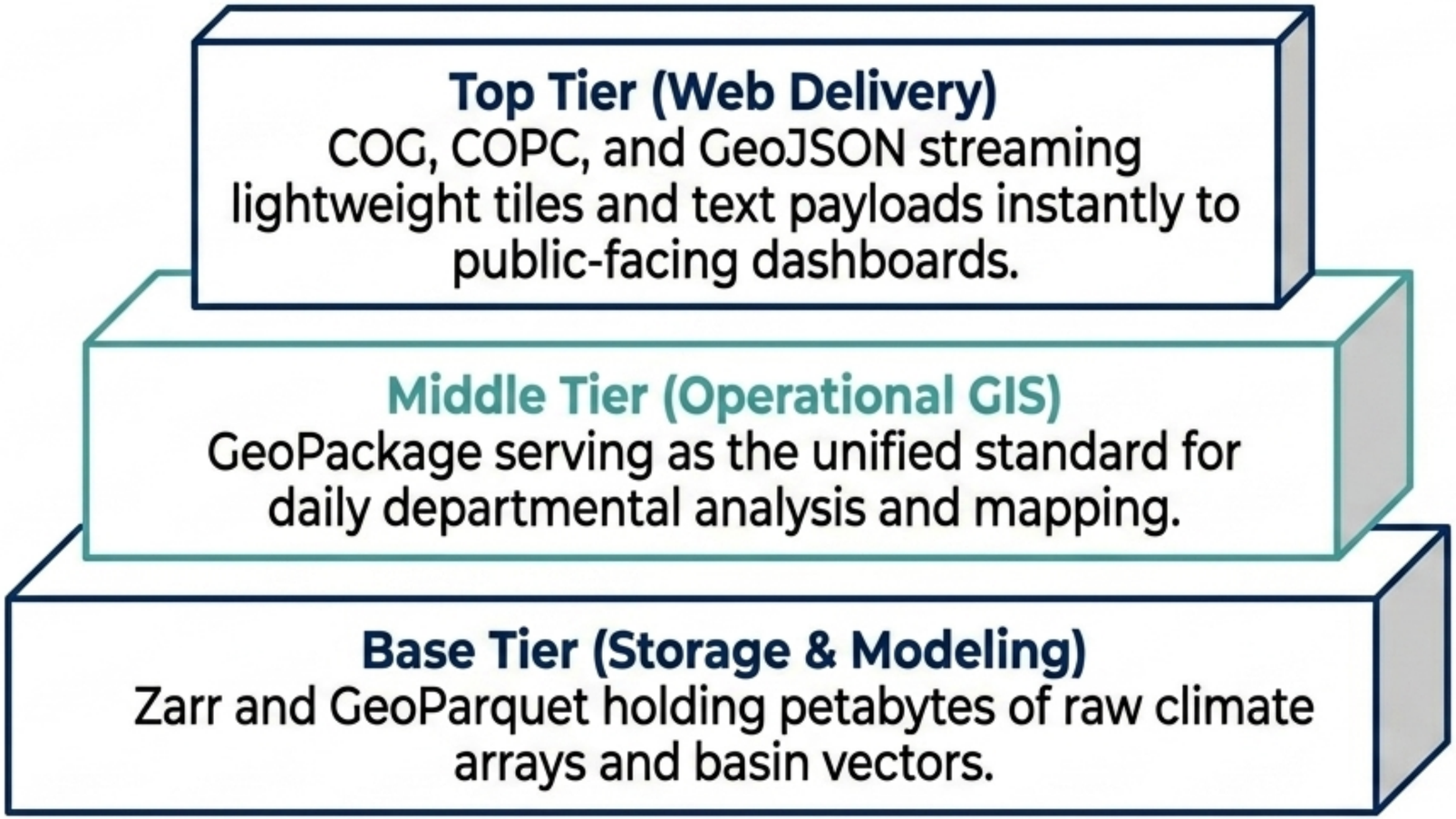
Diagnostic Matrix: The Vector Evolution

Format	Architecture	Scalability Limits	Web Readiness	Primary Government Use Case
Shapefile	Multi-file legacy	2GB limit, 10-char limit	Poor Web	Legacy archival only
GeoPackage	Single SQLite file	**No arbitrary table limits**	Moderate Web	**The modern standard for daily GIS operations**
GeoJSON	Text-based JSON	Small datasets only	**Native Web Standard**	**Lightweight API data exchange**
GeoParquet	Columnar binary	**Petabyte scale** (100x query speed)	High Web	**Massive analytical vector tables**

Diagnostic Matrix: Raster and Big Data Infrastructure

Format	Dimensionality	Cloud Optimization	Primary Hydrological Application
Standard GeoTIFF	2D Matrix	Monolithic (Downloads whole file)	Local raster processing
COG (Cloud-Optimized)	2D Tiled Matrix	HTTP Range Requests (Streams tiles)	Web maps and cloud imagery bases
NetCDF / HDF5	N-Dimensional (X, Y, Z, Time)	Local or cluster processing	Storing decades of time-series climate variables
Zarr	N-Dimensional Chunked	Distributed parallel access	Petabyte-scale concurrent climate modeling (Dask)

Building a Cloud-Native Hydrological Infrastructure



Modernization requires migrating from rigid, localized files to scalable, cloud-optimized formats. Adopting columnar and chunked N-dimensional architectures drastically reduces bandwidth costs while enabling unprecedented computational speed for national hydrological modeling.