Cloud-Optimized OSM GeoParquet Data Service for Switzerland and Beyond

A modular pipeline for converting OpenStreetMap data into cloud-native GeoParquet files

Graduate

Fadil Smajilbasic

Nils-Robin Grob

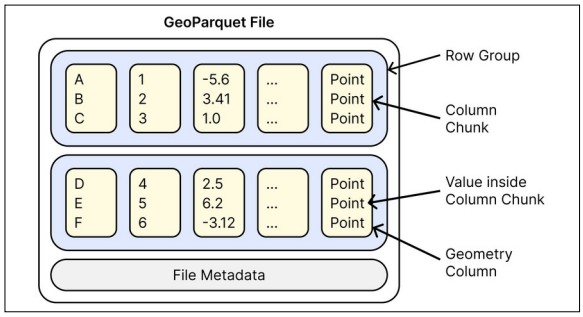
Matthias Hersche

Introduction: OpenStreetMap (OSM) is the most comprehensive openly licensed geospatial vector datasets, containing an estimated 60–90 million points of interest (POIs). While this total is comparable to Overture Maps 61 million, OSM distinguishes itself through crowdsourced data richness and openness. However, OSM's raw data structure, based on a graph of nodes, ways, and relations, combined with a community-driven tagging system, poses challenges for scalable querying and analysis. Modern cloud-native data formats like GeoParquet address some of these challenges by enabling SQL-based access without server infrastructure, but require robust preprocessing. Although Overture Maps aims to provide a unified schema and analysis-ready data, its lack of open-source software and unreliable data sources limit its openness and transparency. This thesis presents an open, reproducible pipeline that transforms country extracts, e.g. from Geofabrik, with an initial focus on the D-A-CH-LI region. The pipeline transforms these extracts into simplified, analysis-ready GeoParquet files aligned with Overture Maps Places and Divisions themes. Transforming OSM data into Overture-aligned themes involves the conversion of the OSM data structures into the layered, tabular formats preferred by geographic information systems.

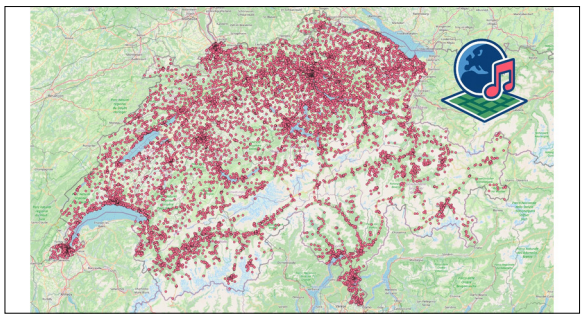
Approach / Technology: The software and data architecture were evaluated specifically for country-scale data extracts to ensure reliability, reproducibility, and scalability. Additionally, multiple spatial partitioning strategies (e.g. KDB-tree, S2) were extensively evaluated, serving the purpose of splitting large 2D data into multiple GeoParquet files, enabling efficient client-side filtering and downstream interoperability. The implemented solution is a modular Extract-Transform-Load (ETL) pipeline built with open-source technologies: osm2pgsql with Lua for data ingestion, PostgreSQL/PostGIS for spatial processing and schema alignment, and Python with DuckDB and PyArrow for conversion into the GeoParquet format. The entire system operates as a CI/CD-enabled DataOps pipeline, leveraging GitLab for orchestration, Docker for containerization, and MinIO for object storage. A prototype for automated vandalism detection and exclusion is used to implement basic data quality assurance.

Result: The result is Cadence Maps: a fully automated OSM GeoParquet data service for the D-A-CH-LI region that can be updated weekly. The data is publicly accessible, hosted on S3-compatible (Minio) storage, accompanied by a static website offering release information and documentation. The service supports efficient client-side querying through tools such as DuckDB and QGIS, eliminating the need for full dataset downloads. A query like: "SELECT names.primary FROM read_parquet ('s3://cadencemaps/release/2025-05-13/theme=places/type=place/country=CH/*', filename=true, hive_partitioning=1) WHERE categories.primary = 'restaurant';" demonstrates selective data extraction using hive-compatible S3 prefixes. The pipeline ingests country-specific extracts from Geofabrik and applies additional KDB-tree based partitioning for large countries to support stable and performant data access. The underlying pipeline adheres to DataOps principles, featuring continuous delivery and automated validation through anomaly detection. Our evaluation confirmed that full dataset updates can be generated in under 24 hours. This provides a robust, reproducible and scalable basis for delivering geospatial data on a global scale.

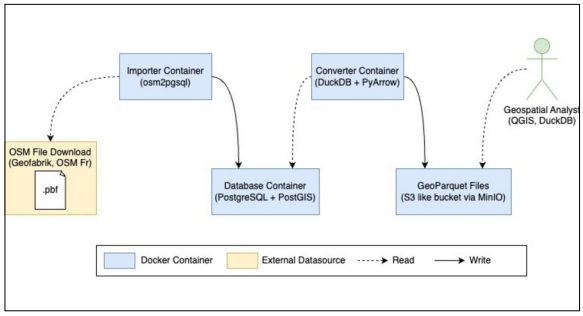
Diagram of a GeoParquet file layout showing row groups, column chunks, and geometry columns with point values.
Own presentment



Map of Switzerland showing all restaurants as red dots accessed via our Cadence Maps service using QGIS.
Map and data (c) OpenStreetMap contributors ODbL



ETL pipeline showing OpenStreetMap import, conversion to GeoParquet files, and access by geospatial analysts.
Own presentment



Advisor
Prof. Stefan F. Keller

Co-Examiner
Claude Eisenhut, Eisenhut Informatik AG, Burgdorf, BE

Subject Area
Software Engineering