BOTTLENECKS TO SURVIVAL
DATABASE REPORT

Prepared by Brahm White-Gluz

May 2024

PACIFIC SALMON FOUNDATION

PACIFIC SALMON FOUNDATION
320 - 1385 W 8TH AVE
VANCOUVER, BC V6H 3V91

Photo credit: Danny Swainson

TABLE OF CONTENTS

Photo credit: Danny Swainson

ABSTRACT

Through the Strait of Georgia Data Centre's partnership with UBC's Institute of Oceans & Fisheries, the Bottlenecks team was able to secure cloud infrastructure to host a Postgres database to house their operations and research data. Furthermore, modern open source software was used to develop an intuitive web-based interface for researchers to interact with the data, upload data from fieldwork, perform ad hoc queries, generate data visualizations, and create dashboards for internal and external reporting. Leveraging state-of-the-art database technologies, data files can be processed automatically as they are inserted to the database through the use of triggers and stored procedures, referential integrity can be enforced through the use of constraints, and performance can be improved through the use of indexes.

SYSTEM ANALYSIS

The Bottlenecks program is an ambitious, multi-year initiative across 13 watersheds on Vancouver Island with multiple community partners. Thousands of PIT tags are being applied in each system each year, along with associated biological information and metadata collected by researchers. This rapid growth in scale and diversity of data therefore required a suitable data management system to be established, to facilitate the complex analyses proposed by this project.

Data are collected via field work, where researchers catch, tag, and release fish and record their observations. Data may include logistical data (e.g. date, crew, hook depth, etc.), biodata (e.g. fork length, weight, clip status, etc.), geospatial data (e.g. coordinates), and genetic information (e.g. tissue sample IDs to be submitted for genetic analyses). These data collections occur in freshwater (in-river), estuarine (beach seine), and marine (microtroll) environments.

Data are also collected in hatcheries, where researchers tag fish prior to their release to the freshwater environment. Information such as fork lengths and clip status may be taken from juveniles that are tagged at hatcheries. Other information collected may be related to the release of these individuals, including the coordinates of their release location, which can be used to inform trends about specific regions. Data are also collected on tagging related mortalities and rejections, to help inform and improve tagging strategies.

Detection data is collected from in-river antenna arrays, which log the out migrations and returns of PIT-tagged fish. The detection data helps shape the understanding of outmigration and return trends among different stock groups over years.

PIT tags can also be redetected during predator or fishery scans. Our biologists target heron rookeries and pinniped haul-outs to scan for tags that may have been consumed by predators and deposited in their nesting areas or haul-outs. These detections provide insight into the predation behaviors that vary across regions and over time. Similarly, biologists scan cleaning tables where recreational fishermen may have caught and cleaned PIT tagged fish. Any discarded tags remaining at cleaning tables may be scanned and provide insight into which stocks are most susceptible to the recreational fishery and the accuracy of our current estimates of recreational fishery impacts.

SYSTEM DESIGN

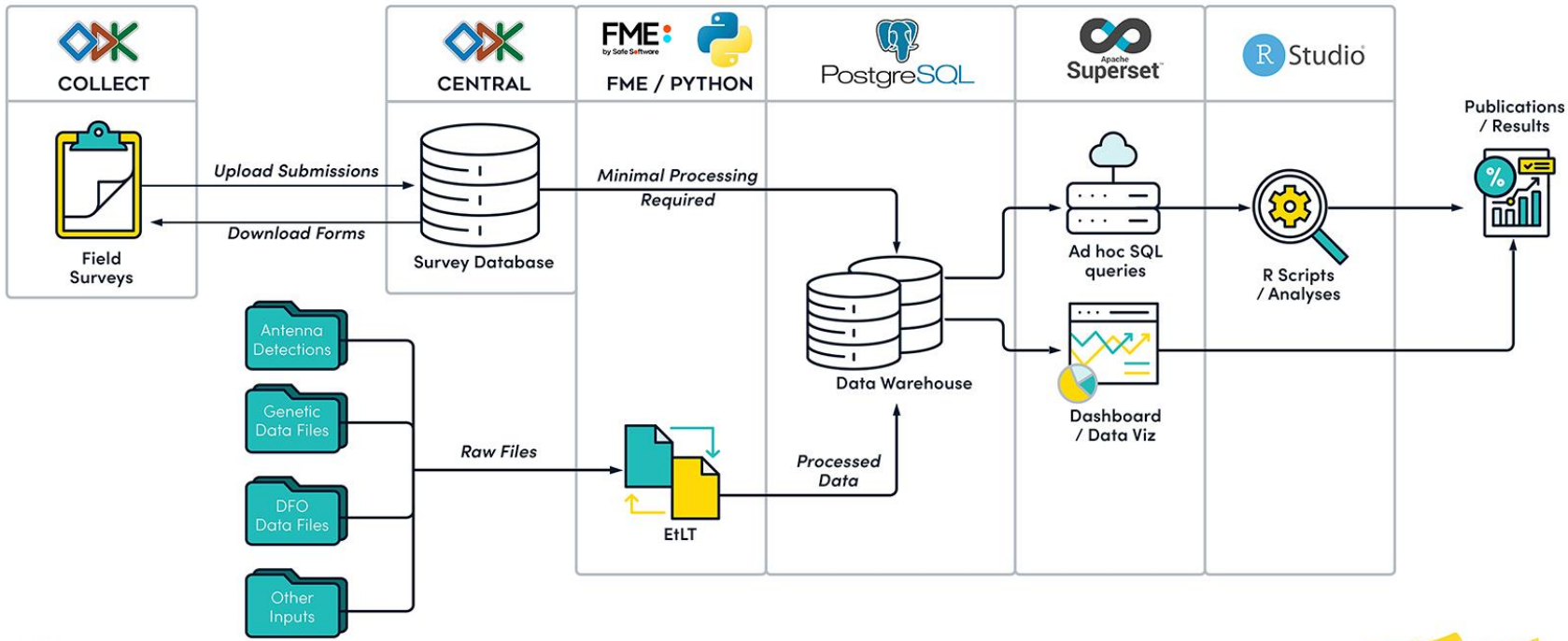
To begin, we developed a data dictionary to help the team standardize and arrive at a consensus on the data definitions for all of the files used by the project. The data dictionary serves as a reference for metadata for all of the files and database tables used by the project.

Significant initial research went into designing a database model for the project, considering analytical purposes and data integrity. A normalized database schema was developed through interviews with team members, reviewing documentation, analyzing historical data, etc. Through the use of constraints, referential integrity can be enforced on the data, helping to ensure the accuracy of the resulting queries performed on those tables. Using triggers, data can be processed automatically as it is uploaded by researchers, thereby standardizing the processes and reducing the risk of human error. Indexes are applied to improve query performance on frequently accessed tables. Views are created for frequent analyses, and where applicable materialized views are employed to reduce loading times.

Figure 1 (following page) - The system design for the Bottlenecks' Information System. This system is an 'end-to-end' solution, including elements for data collection, processing, storage, and analysis.

STRAIT OF GEORGIA DATA CENTRE

BOTTLENECKS – PROCESS FLOW



PACIFIC SALMON FOUNDATION

SYSTEM IMPLEMENTATION

The implementation of the information system for the Bottlenecks project involved multiple software components, described in the architecture diagram in Figure 1. Firstly, cloud infrastructure was provided to the team through the Strait of Georgia Data Centre's established partnership with UBC's Institute of Oceans & Fisheries. Through this partnership, a remote Ubuntu server was set up to host the system. On this server, the various software components were orchestrated using Docker Compose, a solution for managing containerized applications. Fundamentally, the system consists of a Postgres data warehouse and a front-end interface through which team members can interact with the database.

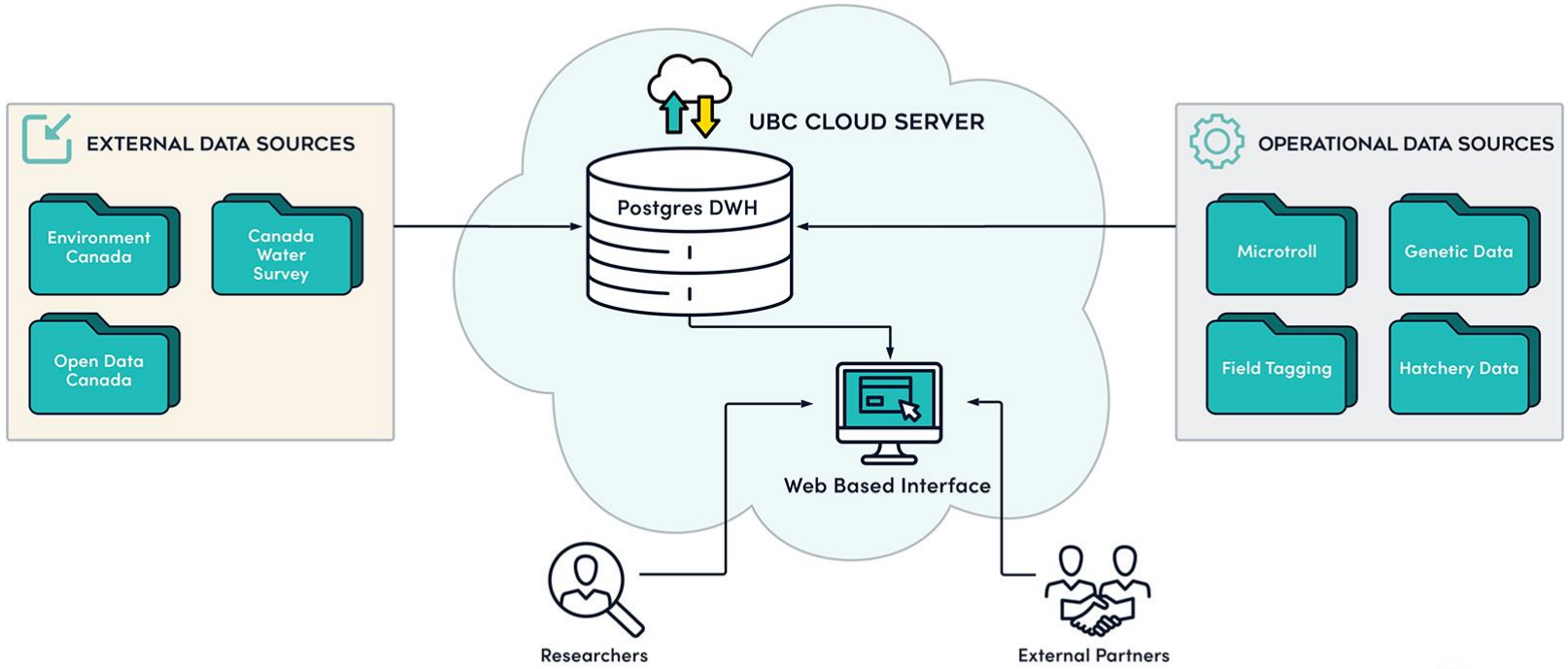
Our proposed system illustrates the different software components and how they relate as part of a larger database system used to manage the research and operational data generated by the Bottlenecks project (Figure 1). To begin with the survey data collected by researchers in the field, we propose using ODK as a centralized survey database. By using a centralized digital survey database, data can be collected by researchers in the field, immediately uploaded to the database and processed to be included in the data warehouse. This process reduces the risk of human error by streamlining the data pipeline, minimizing human involvement. We propose using Python to extract, transform & load (often referred to as ETL) data from the dedicated survey database to the data warehouse, where it will be incorporated into the data model and used in queries to support reporting and analytics. These processes can be scheduled automatically using cron, a Linux daemon for scheduling and automating tasks. Upon insertion to the data warehouse, triggers are activated which further process the data and transmit it to its destination tables. Constraints are applied to the columns of those tables, ensuring that only valid data can be inserted. Views are used to abstract data from their original tables and simplify regular querying.

The development of the information system is ongoing, and some of the proposed components have yet to be fully implemented, such as the ODK survey database. That said, development is very fast-paced, and over the course of just a few months there have been considerable advancements; this trend is expected to continue into the future as new developments are implemented frequently. The next steps include; writing foreign data wrappers to external databases to incorporate those sources into the data warehouse, refining the constraints & triggers applied to the input tables, and further implementation of the normalized data model.

Figure 2 (following page) - Data sources, users and their interactions with the data system. The system context diagram depicted above roughly illustrates the different actors and how they interact with the system, as well as the various input sources of data that flow into the system.

STRAIT OF GEORGIA DATA CENTRE

BOTTLENECKS - NEW SYSTEM



PACIFIC SALMON FOUNDATION

Through the web-based interface, team members can perform ad hoc queries to search for data from the database, generate advanced data visualizations, build interactive dashboards, etc.

Some of the kinds of products which can be developed using this system include:

- Interactive data visualizations, ranging from simple pie charts to advanced geospatial maps.
- Customized dashboards including related visualizations that can be used for internal and external reporting.
- Reproducible analytical queries, saved in a semantic layer for less technical users to understand.

Some of the primary benefits of this system include:

- A centralized repository for data from all research operations, ensuring that all products derived from the database use the same sources of data.
- A semantic layer which enables less technical users to easily interact with the database and understand the datasets they are working with.
- Users can easily explore the data using standard SQL queries, and quickly iterate with a drag-and-drop visualization builder.
- Integrity of the data can be enforced through automated stored procedures & constraints.

Examples of data visualizations can be seen below in Figure 3 and 4.

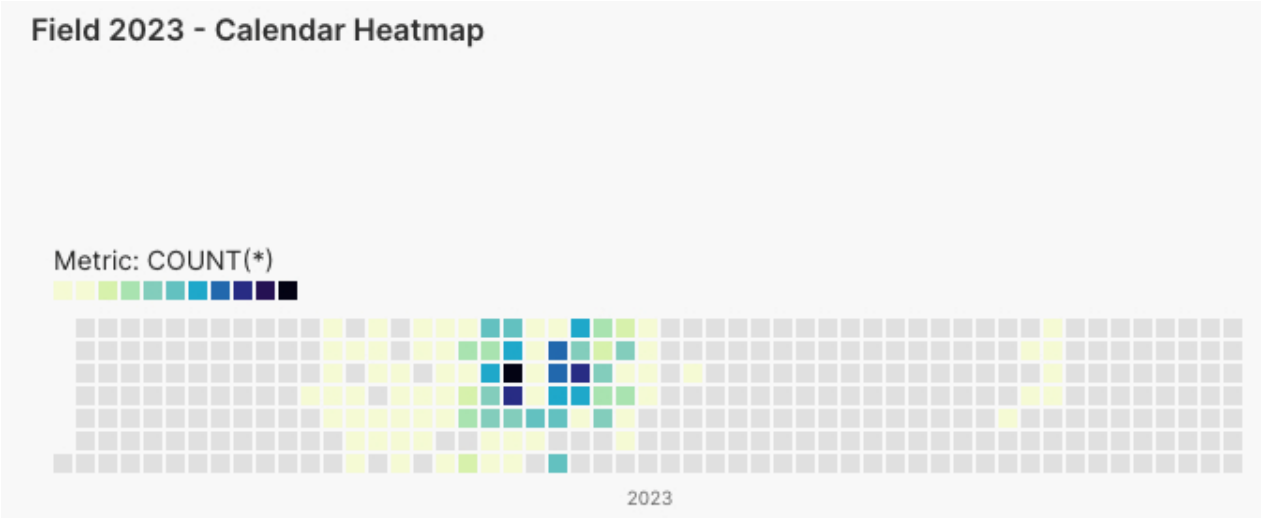


Figure 3 - A calendar heatmap depicting the days of the year spent tagging in the field. Each square represents a day of the year, and the shading of the square indicates the count of the number of tagging events on that day.

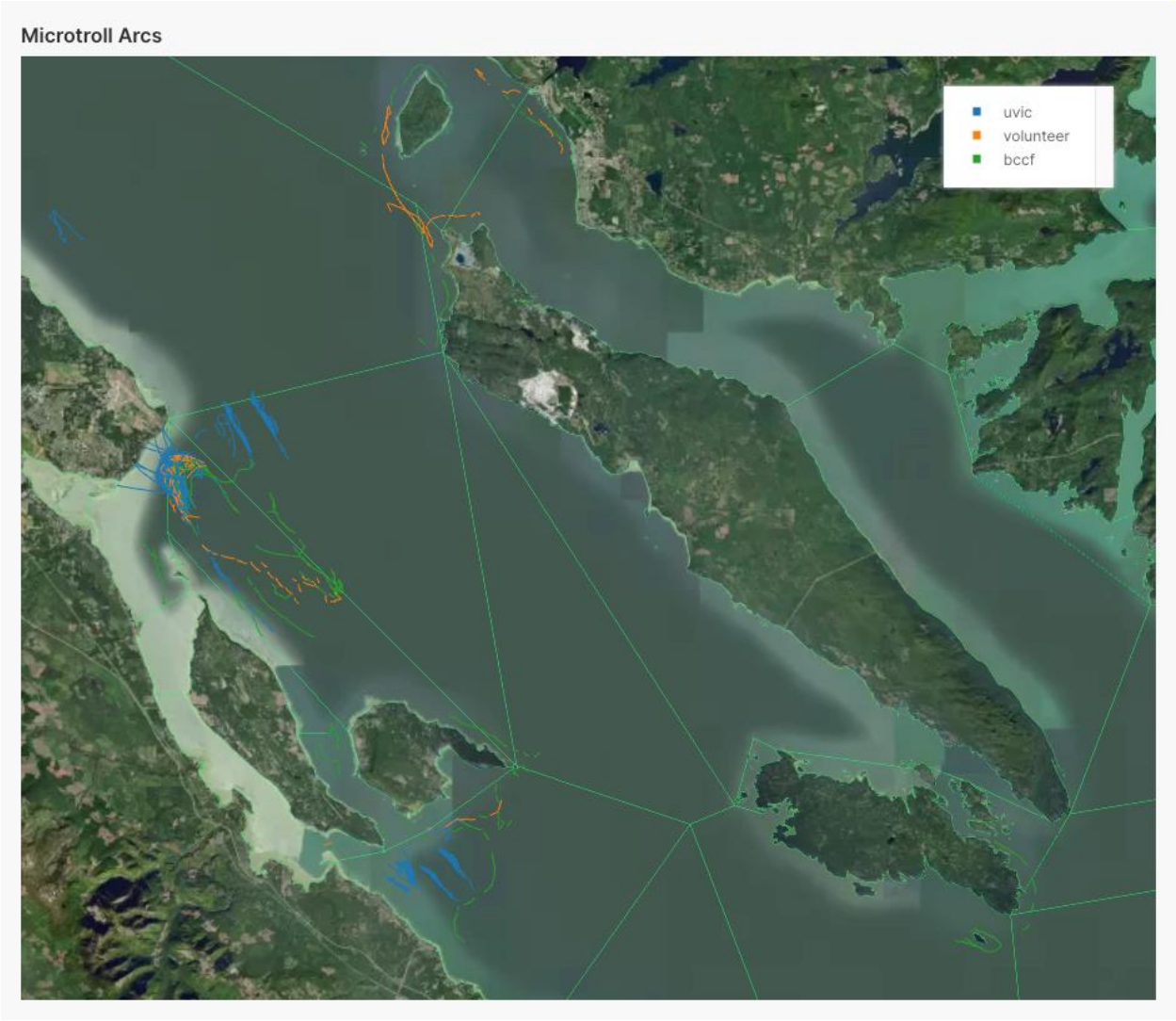


Figure 4 - A map depicting individual microtrolling sets across PFMA subareas in the Strait of Georgia. Colors indicate which crew conducted the microtrolling.

Furthermore, other data sources have begun to be integrated into the data warehouse; these include flow data from the Canada Water Survey, climate data from Environment & Climate Change Canada, etc. The integration of external data sources expands our analytical capabilities beyond what is possible using just the data from our research operations, and opens opportunities to answer fascinating questions beyond the initial scope of the project.

CONCLUSION

The Bottlenecks' Information System is still in its infancy; additional elements to be added to the system include the open source data collection system ODK, and foreign data wrappers to connect with the Strait of Georgia Data Centre's existing repository of geospatial data. With those additions, the information system will integrate data collection directly from the source to the database for immediate use, and be able to incorporate the wide array of data housed at the SGDC into sophisticated analyses. In addition to these promising opportunities for growth, a partnership has been established with UBC's Masters of Data Science program, and in May 2024 the team welcomed 4 graduate students to work with us for 8 weeks as part of their capstone project.

The design & implementation of the Bottlenecks project's information system addresses many of the issues surrounding data management for a broad, multi-year research project involving disparate sources of data. The use of modern software engineering techniques allows the team to leverage state-of-the-art data management technologies to help ensure the accuracy and reliability of the data collected throughout the project. The information system that has been developed for the Bottlenecks project not only ensures reproducible ongoing research operations, but also opens opportunities for further incorporating other sources of data to address larger overarching questions as the project continues to grow and evolve over time.

Photo credit: Eiko Jones

Fisheries and Oceans Canada | Pêches et Océans Canada

BRITISH COLUMBIA

Funding for this project is provided by the BC Salmon Restoration and Innovation Fund, a contribution program funded jointly between Fisheries and Oceans Canada and the Province of BC.

PACIFIC SALMON FOUNDATION