PROJECT 2 REPORT

GROUP 07
Annaliisa Vask, Remi Raugme, Selene Margaret Pruuden, Victoria Prins

Our GitHub: https://github.com/Lollover45/data-engineering-project

A short recap of our project to better understand the topic and, therefore, our challenges compared to project 1. Our aim is to analyze U.S. data on bee occurrences and on diseases and pests, which represent a major issue for multiple stakeholders, including beekeepers (who rely on healthy colonies for honey production and pollination services), policy makers (who develop regulations and conservation policies), and biologists (who study bee population dynamics and ecosystem health). Understanding seasonal health stressor patterns helps beekeepers optimize treatments and reduce colony losses.

1. FEEDBACK Q&A

Some of the questions mentioned in the last feedback deserve further explanation.

Why is the grain of your Dim.Date table set at such a high level, with the month as the smallest grain?
→ Since most colonies overwinter in a semi-dormant state and become active again in the spring, it is crucial to monitor their health over these seasonal cycles - especially because parasites and diseases can cause entire colonies to collapse during the winter, a risk that beekeepers are keen to avoid. Bee statistics do not change hourly, daily, or even weekly, but rather on a monthly or seasonal basis. The observations are recorded monthly, and there is no benefit in measuring more frequently. Therefore, the grain of our Dim.Date table is set to the month level.

Why does your Dim.Location table use SCD type 1 when longitude and latitude are included, which could easily change for counties, as this is your smallest grain?
→ The coordinates do not represent a county's location but rather indicate the location of individual observations. Since neither of the datasets includes coordinates for each observation, we decided to perform our analysis at the county level. Therefore, the Dim.Location table uses SCD Type 1, as county-level data does not require historical tracking of changes in coordinates.

Why did you choose to use only file ingestion?
→ The datasets we selected come from real-life biological data repositories that only provide file-based ingestion. We chose to embrace the challenges faced by real biologists and aim to develop a robust data pipeline that reflects the practical limitations of the field.

2. BUSINESS BRIEF

2.1 Business Questions

We refined our questions based on the feedback from Project 1 and in alignment with the updates in the star schema. The questions now focus on investigating the relationships between the two datasets.

1. **Question about viruses:**
    a. From the three counties with the highest average virus prevalence - how many bees were detected during the year 2024?
2. **Questions about mites:**
    a. How many bee occurrences are there in the five counties with the fewest Varroa mites?
    b. How many bee occurrences are there in the five counties with the most Varroa mites?
3. **Questions about fungus:**
    a. How many bee occurrences are there in the five counties with the least Nosema fungus?
    b. How many bee occurrences are there in the five counties with the most Nosema fungus?
4. **Question about bees' safety per county:**
    a. Which county is most popular for beekeeping and which is most safe from pests?

3 DATASETS AND ARCHITECTURE

3.1 Datasets

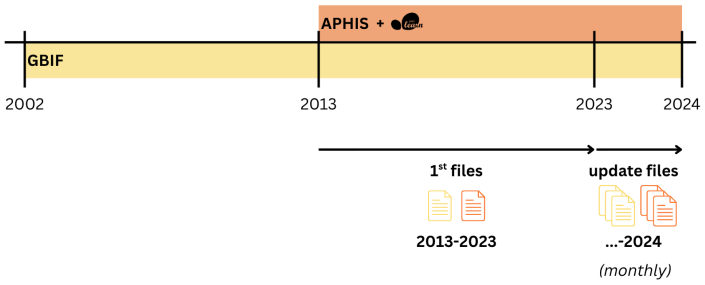

Figure 2. Overview of GBIF and APHIS bee data coverage and ingestion timeline

As originally planned, we combined two datasets that together reflect the overall health of bee populations.

1. The first one (from GBIF) is a detailed and accurate dataset that provides precise information about the taxonomy and distribution of bees across U.S. counties. Data on bees has been added to the repository since 2002. We use only the data related to honey bees.
2. The second one (from APHIS) offers an overview of varroa mites, viruses, and fungal diseases found in bees.

Since the dataset from APHIS covers only the period from June 2013 to December 2024 (Figure 2), we will filter the other dataset from GBIF to match this time range during our data ingestion process. To test the dataset update process, we will first ingest data for the period from June 2023 to December 2023, and then simulate monthly updates by creating smaller files from the year 2024. Additionally, because the dataset from APHIS contains some missing data, we have decided to generate additional observations using the *scikit-learn* package.

3.2 Data Architecture (new digitized version of the diagram)

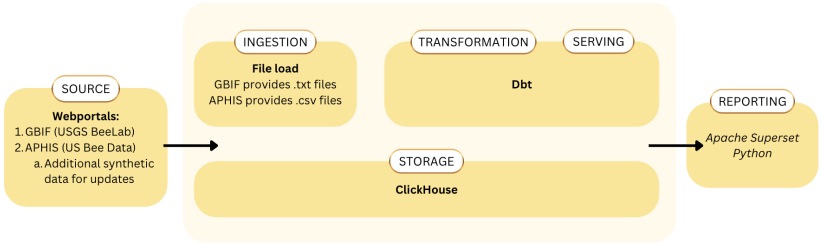

Figure 1. Diagram of Data Flow

4. DATA MODEL

We have updated our star schema based on the feedback. It indicated that it wasn't possible to query information about both datasets simultaneously without many joins. To address this, one could add more Occurrence attributes or create views for cross-organism analysis. However, to keep the schema simple and analytical queries transparent, we decided not to use views yet but to acknowledge them as a possible future solution. We also included some new fields derived from the full date and county. The new fields in Dim.Date are season and decade, and in Dim.Location we added state and country. Our database uses SCD Type 0, but for Dim.Organism we changed it to SCD Type 2 to allow for potential changes in taxonomy or units. Therefore, we added ValidFrom and ValidTo to Dim.Organism.

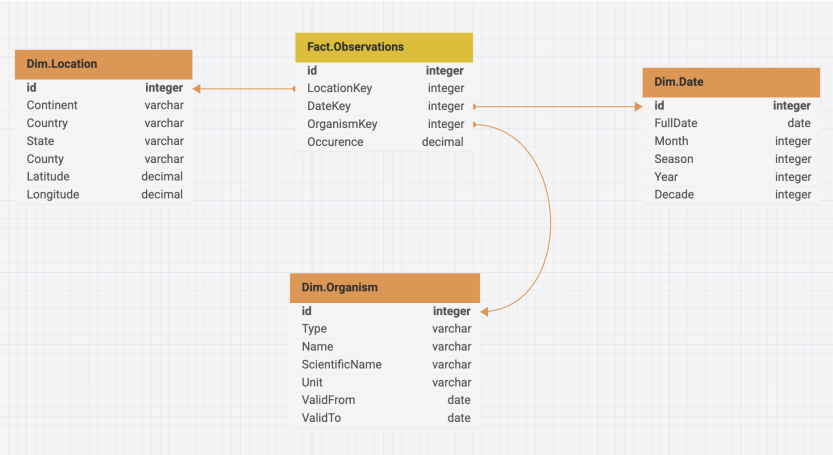

Figure 3. Star schema suitable for querying different types of organism occurrences simultaneously

5. GROUP MEMBERS' ROLES AND CONTRIBUTION

All members attended multiple meetings per week and worked on fixing the issues from project 1. In addition, everyone contributed to debugging the remaining issues that were identified. Furthermore,

- **Annaliisa 15%** prepared the ingestion files from APHIS, generated additional data where necessary, and proposed new analytical SQL queries for our business questions. For this project, she contributed a bit less than others due to health reasons, which the rest of the team fully understood.
- **Victoria 25%** prepared the ingestion files from GBIF and created SQL queries for building both the bronze and gold layers.
- **Selene 25%** led the meetings and planned the teamwork, identified weaknesses in our pipeline design, consulted with the course lecturer and instructors to clarify our questions, created diagrams, and wrote the report.
- **Remi 35%** did the majority of the work on building and maintaining the pipeline. He invested a lot of time into debugging various technical issues, and therefore the team decided to reward him with a higher contribution percentage.

6. LLM DISCLOSURE

https://chatgpt.com/share/6903b734-2cf8-8013-924c-616ff309967e
https://chatgpt.com/share/69074c42-f91c-8002-a826-2fe9cd85b67a
https://chatgpt.com/share/6907a174-6324-8002-a99b-3e507afc418e
https://chatgpt.com/share/6907a22f-8440-8002-8edf-a69944f95d0d
https://chatgpt.com/share/6907a253-5730-8002-8a3e-6f015e404401
https://chatgpt.com/share/6907a262-b848-8002-b4d3-644b393d7d1b
https://chatgpt.com/share/6907a26f-6a10-8002-a5a2-ba01c5569b95