Integrating Community resources into Grapedia's database with Airbyte

Victor García-Carpintero
IBMCP-CSIC
vicgarb4@upvnet.upv.es

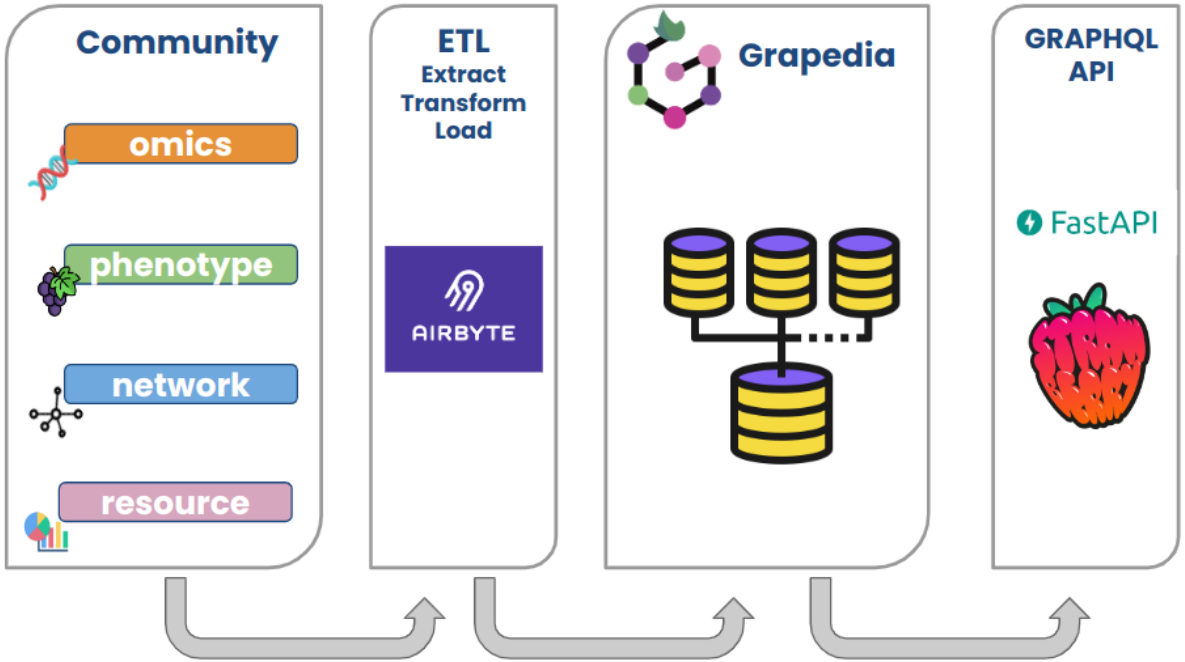
Background

Grapevine scientific community has generated over the past years abundant, high quality resources and knowledge about grapevine: genetics, omics, phenotyping…

- INTEGRAPE has contributed with the standardization of this resources following the principles of findability, accessibility, interoperability, and reusability (F.A.I.R)

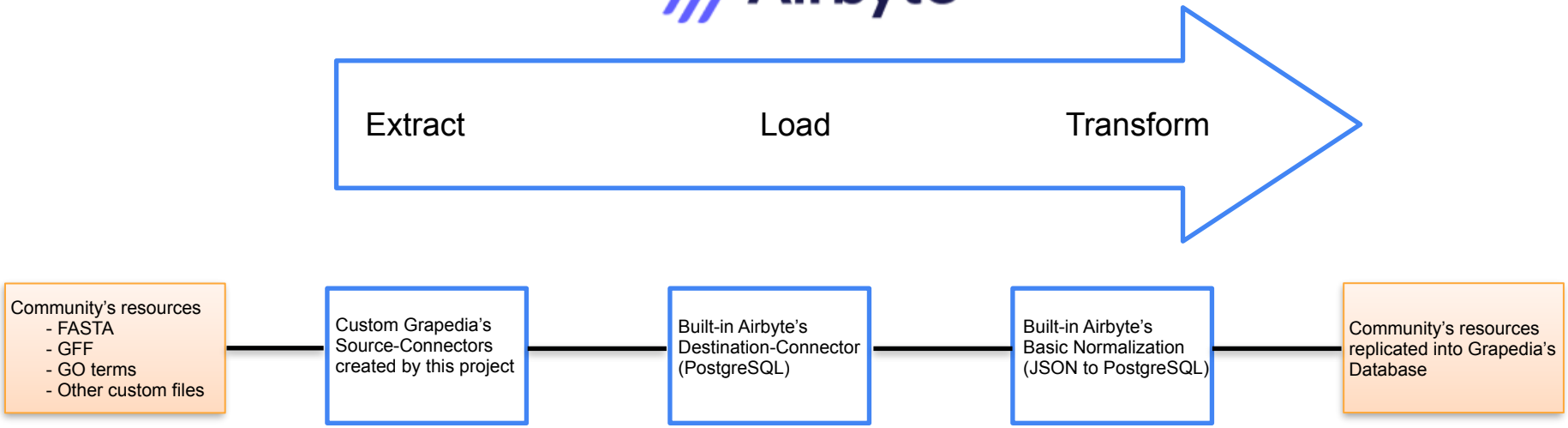
- Grapevine resources remains scattered, though. One of the Grapedia's main goals is to centralize all these resources on an open portal among other services. Integrate all data shared by the community into a federal database is required in order to achieve this.

Airbyte



Source: Marco Moretto

Airbyte



Following Airbyte's EL(T) process, shared Community's resources can be replicated into Grapedia's unified database in a seamless way, regardless of their origin or format.

What is a connector?

A connector is a Docker image built inside Airbyte with the instructions and specifications required to extract or load data.

- Airbyte provides its own connector's core generator so you don't have to do it from scratch.
- Connectors can be coded using java or python (or using Airbyte's web interface).
- Both Source and Destination connectors need to define how users are going to interact with the data stored. These definitions are called streams.
- Connectors can be configured to overwrite existing records o do incremental upgrades into the database.

Connectors development

For this project, connectors were developed to integrate the following data:

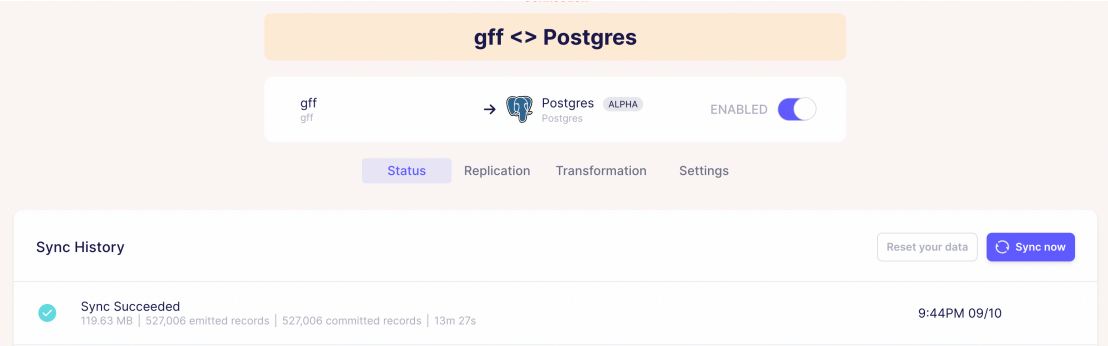
- FASTA file format
- GFF file format
- GO Terms file format
- INTEGRAPE reference gene catalog
- OneGenE output file format

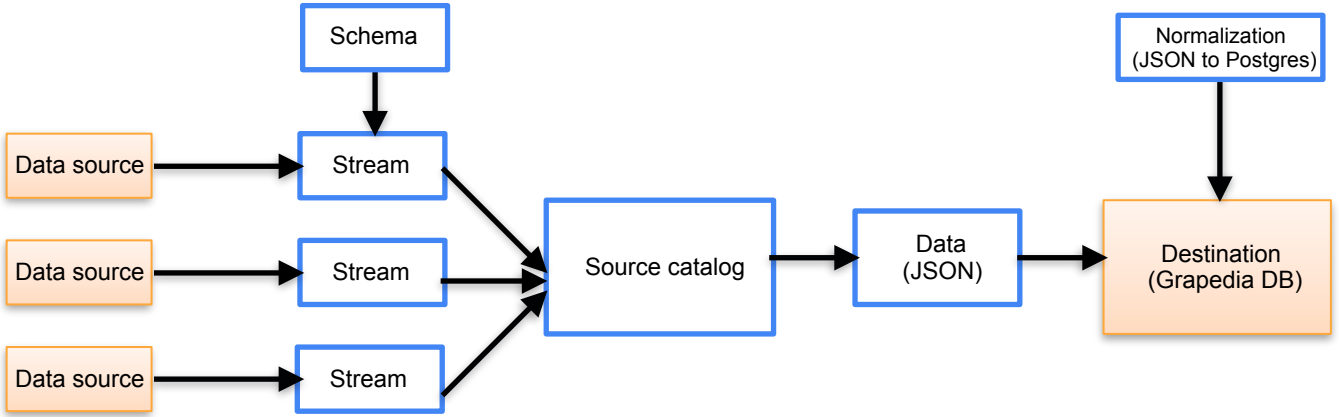
Grapedia connectors' development had two ways of quality assessment:

- Airbyte's checks: Airbyte can check if connectors can communicate with the data storage and check if streams are properly configured with shell commands.

- All connectors were tested with sample data to check if ELT process was working as intended.

Data synchronization





Some considerations

Airbyte provides a simple solution to integrate Community Resources into Grapedia's database while keeping the ability to update previous data or adding more, but:

- Speed performance could be a problem for some volumes of data if constant updates are expected: loading GFF or GO Terms files takes 10 minutes approximately in a laptop, but OneGenE data can take more than 30 hours.

- Airbyte stores all records in raw data even if these records are going to be normalized and needs to be removed manually (as far a I know). This can consume quickly the space available if a proper database maintenance is not done.

- Choosing a proper database system where our data is going to be integrated should be a major concern. For example, adding a GFF file adds more 500 thousand records into a database. If more data is added in the future this could slow database queries and degrade the user's experience. Non-relational database systems should be considered if this is the case.

Acknowledgments

https://www.sequentiabiotech.com/

Marco Moretto
Fondazione Edmund Mach (FEM, Italy)
marco.moretto@fmach.it