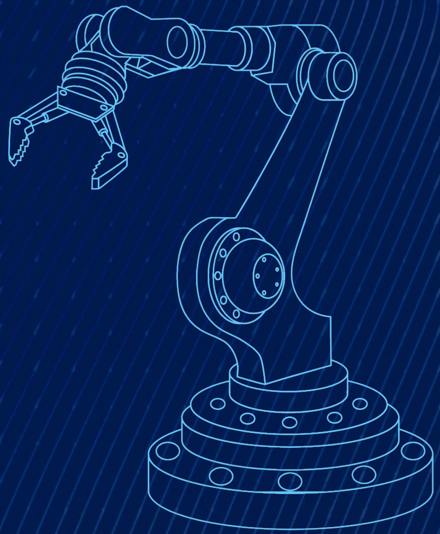
ENVS 5726 Guest Lecture

Building Data Pipelines to Support Mineral Exploration Activity

Abdel Alfahham
alfahham@sas.upenn.edu

Founding Data Engineer

TABLE OF CONTENTS



01

Introduction

Summary and Objective

Career Timeline

- BS Environmental Sciences [2009 - 2013]
- Environmental Consulting [2013 - 2018]
- Penn MSAG [2016 - 2019]
- Data Engineer [2019 - Current]

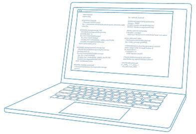
- Self Taught Python (coding in general)
- Intro to SQL in Environmental Consulting
- Intro to GIS/QGIS/CAD in Environmental Consulting

Founding Data Engineer
ecue ai · Full-time
Jul 2024 - Present · 1 yr 6 mos
New York, New York, United States · Remote

Software Engineer
Grata · Full-time
Sep 2022 - Aug 2024 · 2 yrs
New York, New York, United States · Remote

KoBold Metals
Full-time · 2 yrs 11 mos
San Francisco Bay Area · Remote

Data Engineer
Nov 2019 - Sep 2022 · 2 yrs 11 mos

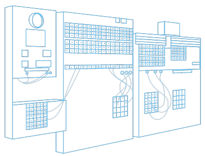


Lecture Objective

1. Provide a *detailed but practical overview* on the various components of an end-to-end *data platform*.

2. Introduce critical *data engineering, data infrastructure and software development concepts.*

3. Demonstrate the *impact of data engineering in earth science.*



Background

- During this lecture I will be discussing and demonstrating a hypothetical pipeline that I created to share the knowledge I have learned from the past 7 years or so as a data engineer.

- The main objective is to share the fundamentals and give jumping off points for individuals that are interested in pursuing data engineering or data science roles.

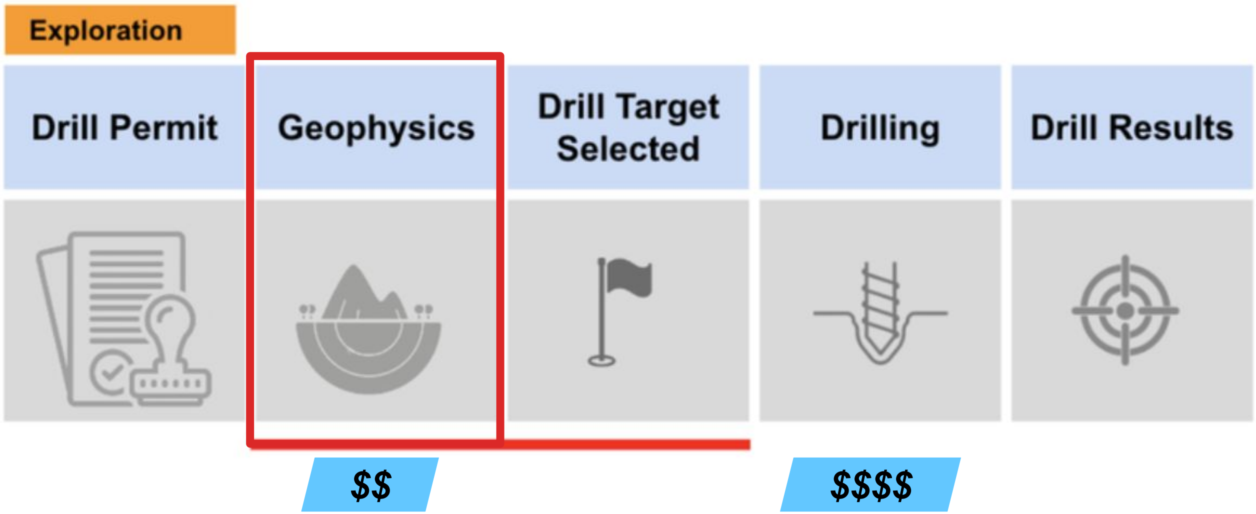
02

Geophysical Surveys

Ground EM Fundamentals

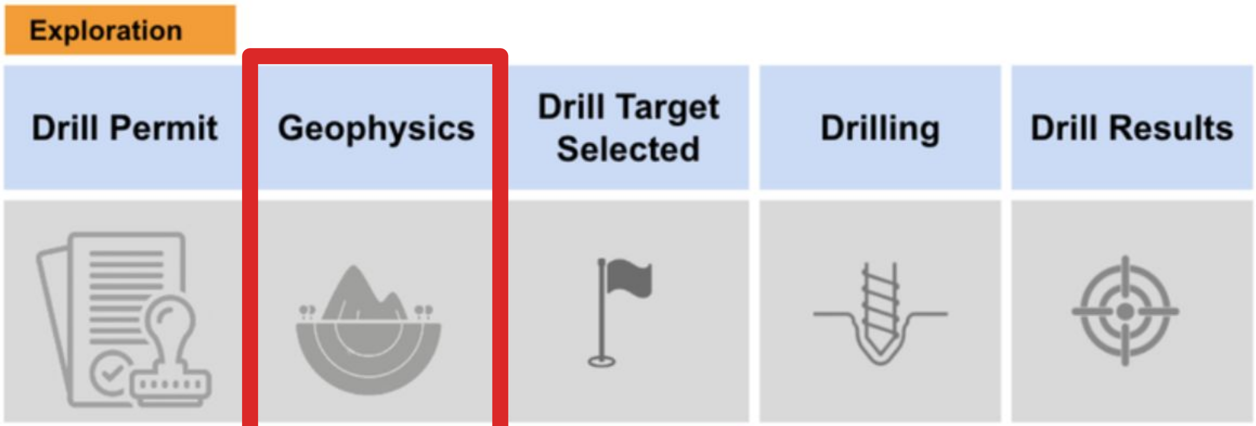
Geophysics in Mineral Exploration

ENVS 5726



Geophysical surveys allows exploration teams to identify promising areas without having to dig or drill, saving time and money in the early stages of mineral exploration.

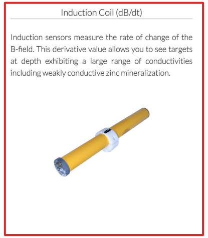
Types of Geophysical Surveys

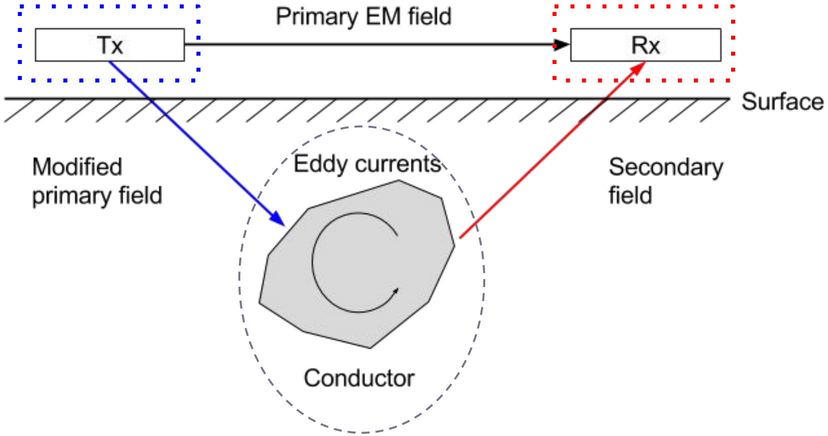


- **Electromagnetics (EM)**
- **Gravity**
- **Induced Polarization (IP)**
- **Magnetics**

Electromagnetic (EM) Surveys

1. An artificial electromagnetic signal from the **transmitter (Tx)** is generated and it penetrates deep into the earth.
2. When the signal hits **conductors** (conductive layers of rock or minerals like nickel, copper, and other base metals) it changes.
3. A **receiver (Rx)** picks up these changed signals and the information is saved.





Electromagnetic (EM) Surveys

What it measures:

- Ground EM surveys detect how well electrical current flows through subsurface materials.
- Highly conductive materials (like ore deposits) conduct electricity easily, while resistive materials don't

How it works:

- An artificial electromagnetic signal is transmitted into the ground, and the response is measured; conductive materials create stronger secondary signals that the receiver detects

Why it matters:

- Conductivity patterns reveal the location, size, and type of subsurface targets; anomalies indicate potential mineral deposits, geological structures.

Revisiting Electromagnetics (EM)

- **Strong Conductors and Anomalies:** Conductors and anomalies are prominent features on an EM survey, these are zones that conduct electricity particularly well. *Note: depending on the type of mineral the company is searching for, they may be looking for low conductivity.*
- **Shape and Depth of Anomalies:** The shape and depth of a conductive anomaly can help with modelling structures underground.
- **Context of Geological Features:** EM survey results are often used in conjunction with other geophysical data such as rock chip sampling, this allows an explorer to be further informed about the potential. For example, an anomaly along a well-known deposit belt or occurrence can serve to increase the confidence of prospectivity in a project.
- **Comparison with previous data: I**f there is historical data and older survey results for the region, it can be compared with the new EM data to give the company a better understanding of its project.

Electromagnetic (EM) Survey Timeline

Conventional Survey Analysis Pace (12 Months)

- **Scoping** — ~3 months
- **Permits** — ~2 months
- **Field Survey** — 3 months
- **Manual Data Processing** — 1 month
- **Analysis** — 2 months

Electromagnetic (EM) Survey Timeline

Proposed Survey Analysis Timeline (Fast Iteration)

Scoping
~3 months

Permits
~2 months

Field Survey
3 months

Auto-Data Processing
~6 hours

ML + Analysis
~6 hours

Electromagnetic (EM) Survey Timeline

ENVS 5726

Proposed Survey Analysis Timeline (Fast Iteration)

Scoping
~3 months

Permits
~2 months

Field Survey
3 months

Auto-Data Processing
~6 hours

ML + Analysis
~6 hours

How would you do it?

???

???

???

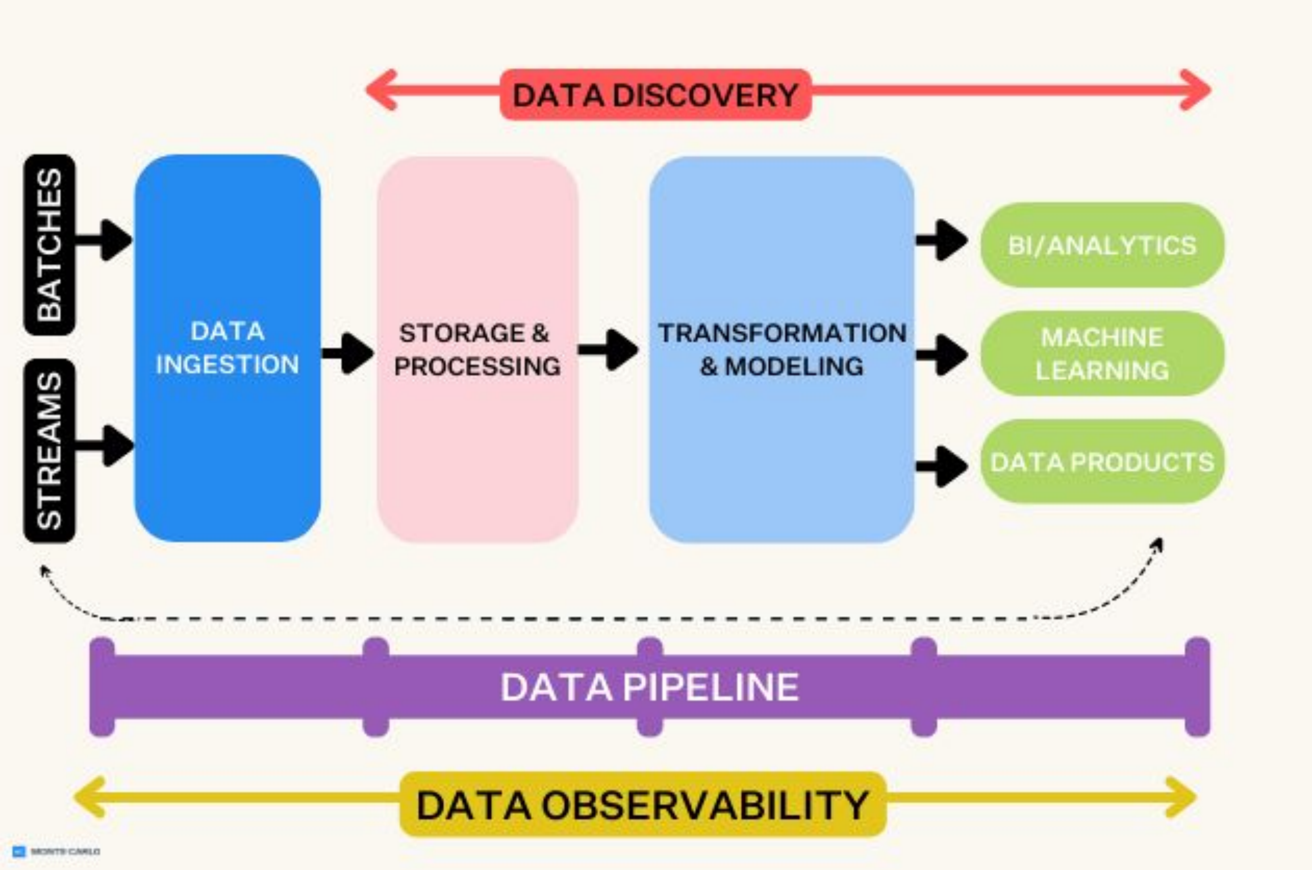
???

03

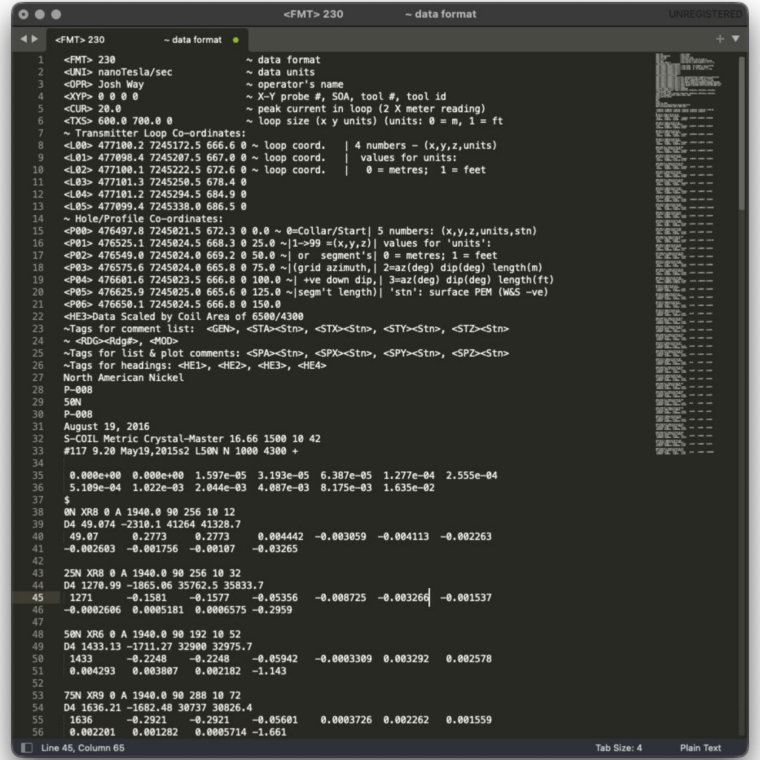
Data Pipelines

Automation, Scaling, Design

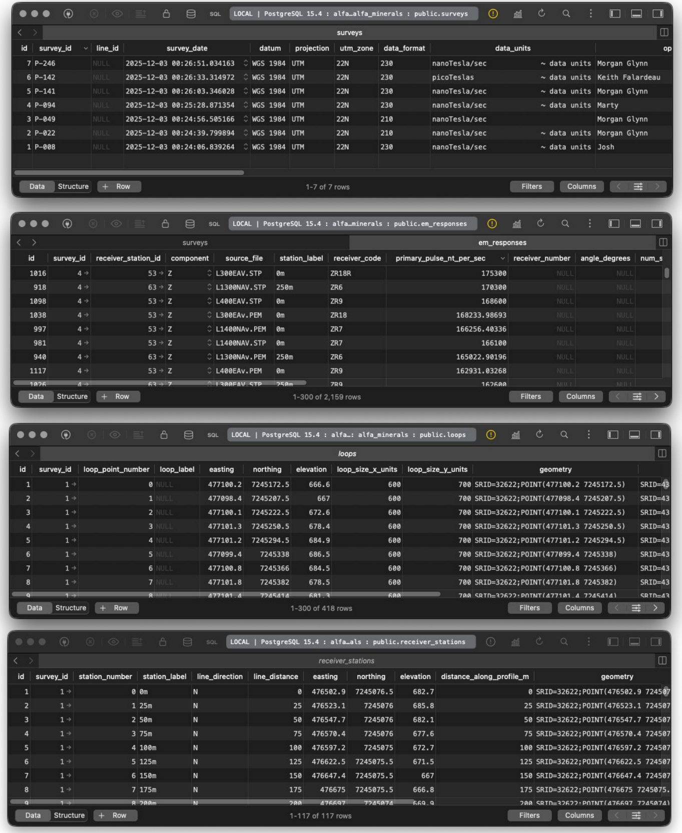
Basic Components of Data Platform



Let’s Start the Demo Now!

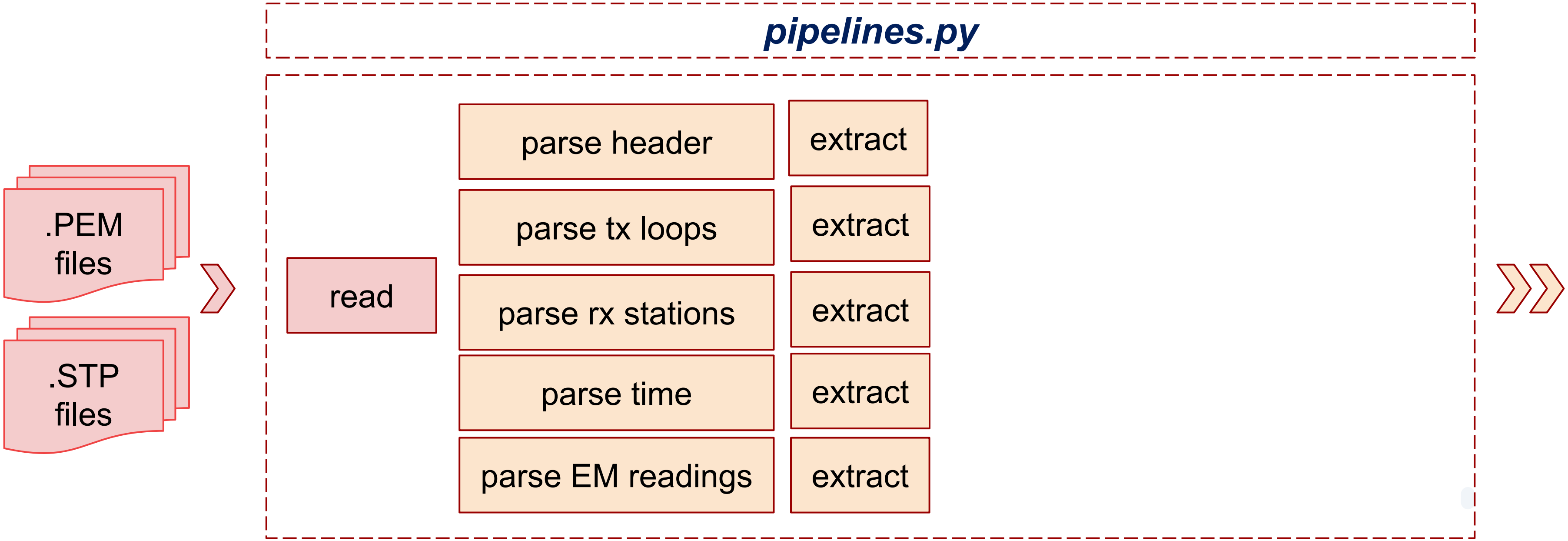


```
<FMT> 230                         ~ data format
<UNI> nanoTesla/sec               ~ data units
<OPR> Josh Way                    ~ operator's name
<XYP> 0 0 0 0                     ~ X-Y probe #, SOA, tool #, tool id
<CUR> 20.0                        ~ peak current in loop (2 X meter reading)
<TXS> 600.0 700.0 0               ~ loop size (x y units) (units: 0 = m, 1 = ft
~ Transmitter Loop Co-ordinates:
<L00> 477100.2 7245172.5 666.6 0 ~ loop coord.   | 4 numbers - (x,y,z,units)
<L01> 477098.4 7245207.5 667.0 0 ~ loop coord.   |  values for units:
<L02> 477100.1 7245222.5 672.6 0 ~ loop coord.   |   0 = metres;  1 = feet
<L03> 477101.3 7245250.5 678.4 0
<L04> 477101.2 7245294.5 684.9 0
<L05> 477099.4 7245338.0 686.5 0
~ Hole/Profile Co-ordinates:
<P00> 476497.8 7245021.5 672.3 0 0.0 ~ 0=Collar/Start| 5 numbers: (x,y,z,units,stn)
<P01> 476525.1 7245024.5 668.3 0 25.0 ~|1->99 =(x,y,z)| values for 'units':
<P02> 476549.0 7245024.0 669.2 0 50.0 ~| or  segment's| 0 = metres; 1 = feet
<P03> 476575.6 7245024.0 665.8 0 75.0 ~|(grid azimuth,| 2=az(deg) dip(deg) length(m)
<P04> 476601.6 7245023.5 666.8 0 100.0 ~| +ve down dip,| 3=az(deg) dip(deg) length(ft)
<P05> 476625.9 7245025.0 665.6 0 125.0 ~|segm't length)| 'stn': surface PEM (W&S -ve)
<P06> 476650.1 7245024.5 666.8 0 150.0
<HE3>Data Scaled by Coil Area of 6500/4300
~Tags for comment list:  <GEN>, <STA><Stn>, <STX><Stn>, <STY><Stn>, <STZ><Stn>
~ <RDG><Rdg#>, <MOD>
~Tags for list & plot comments: <SPA><Stn>, <SPX><Stn>, <SPY><Stn>, <SPZ><Stn>
~Tags for headings: <HE1>, <HE2>, <HE3>, <HE4>
North American Nickel
P-008
50N
P-008
August 19, 2016
S-COIL Metric Crystal-Master 16.66 1500 10 42
#117 9.20 May19,2015s2 L50N N 1000 4300 +

 0.000e+00  0.000e+00  1.597e-05  3.193e-05  6.387e-05  1.277e-04  2.555e-04
 5.109e-04  1.022e-03  2.044e-03  4.087e-03  8.175e-03  1.635e-02
$
0N XR8 0 A 1940.0 90 256 10 12
D4 49.074 -2310.1 41264 41328.7
 49.07      0.2773     0.2773     0.004442  -0.003059  -0.004113  -0.002263
-0.002603  -0.001756  -0.00107   -0.03265

25N XR8 0 A 1940.0 90 256 10 32
D4 1270.99 -1865.06 35762.5 35833.7
 1271      -0.1581    -0.1577    -0.05356   -0.008725  -0.003266  -0.001537
-0.0002606  0.0005181  0.0006575 -0.2959

50N XR6 0 A 1940.0 90 192 10 52
D4 1433.13 -1711.27 32900 32975.7
 1433      -0.2248    -0.2248    -0.05942   -0.0003309  0.003292   0.002578
 0.004293   0.003807   0.002182  -1.143

75N XR9 0 A 1940.0 90 288 10 72
D4 1636.21 -1682.48 30737 30826.4
 1636      -0.2921    -0.2921    -0.05601    0.0003726  0.002262   0.001559
 0.002201   0.001282   0.0005714 -1.661
```

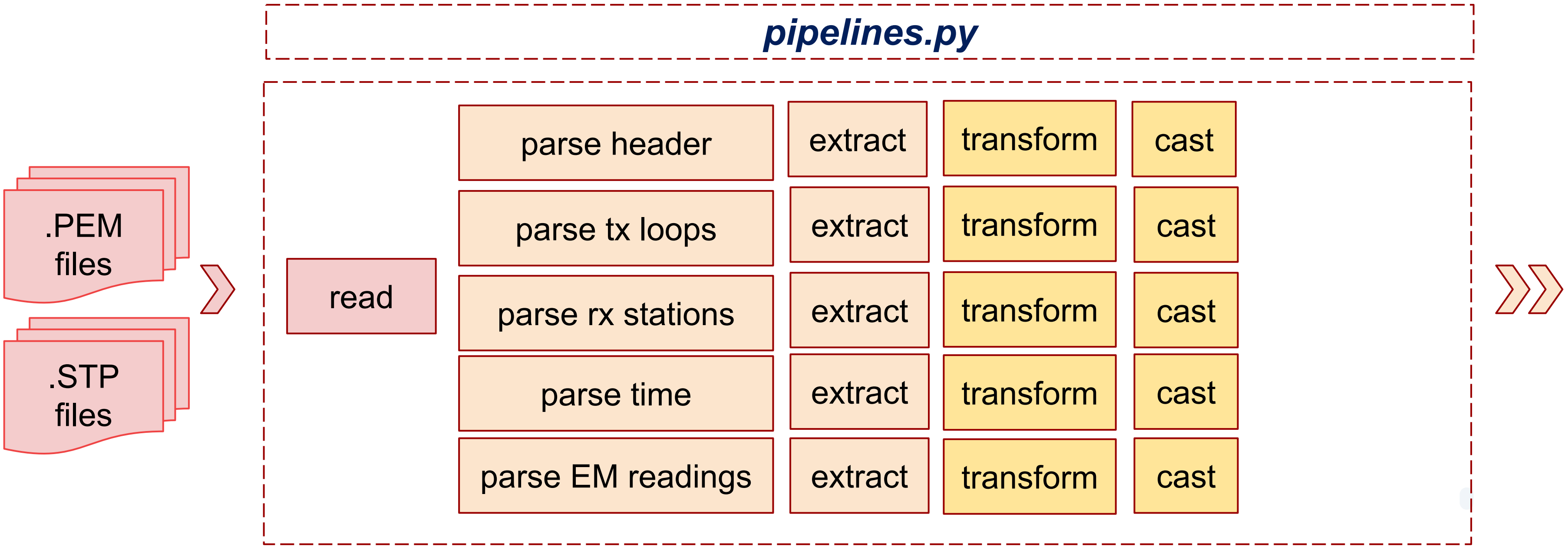



Basic Components of Data Pipeline

ENVS 5726

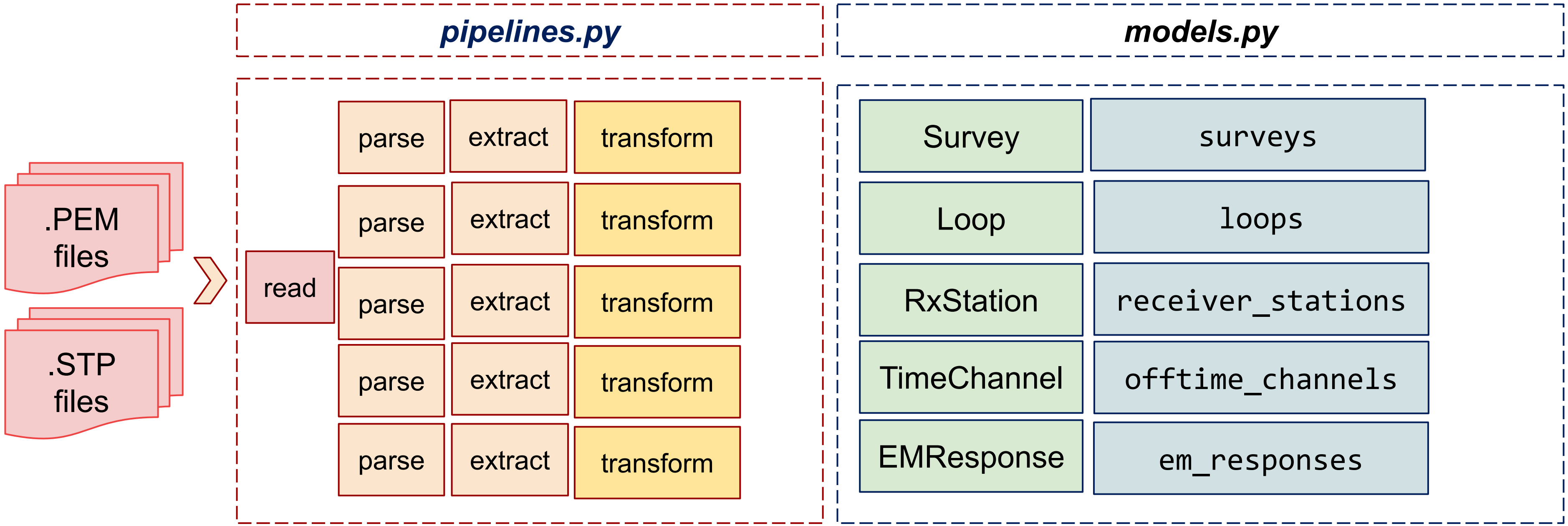


Basic Components of Data Pipeline

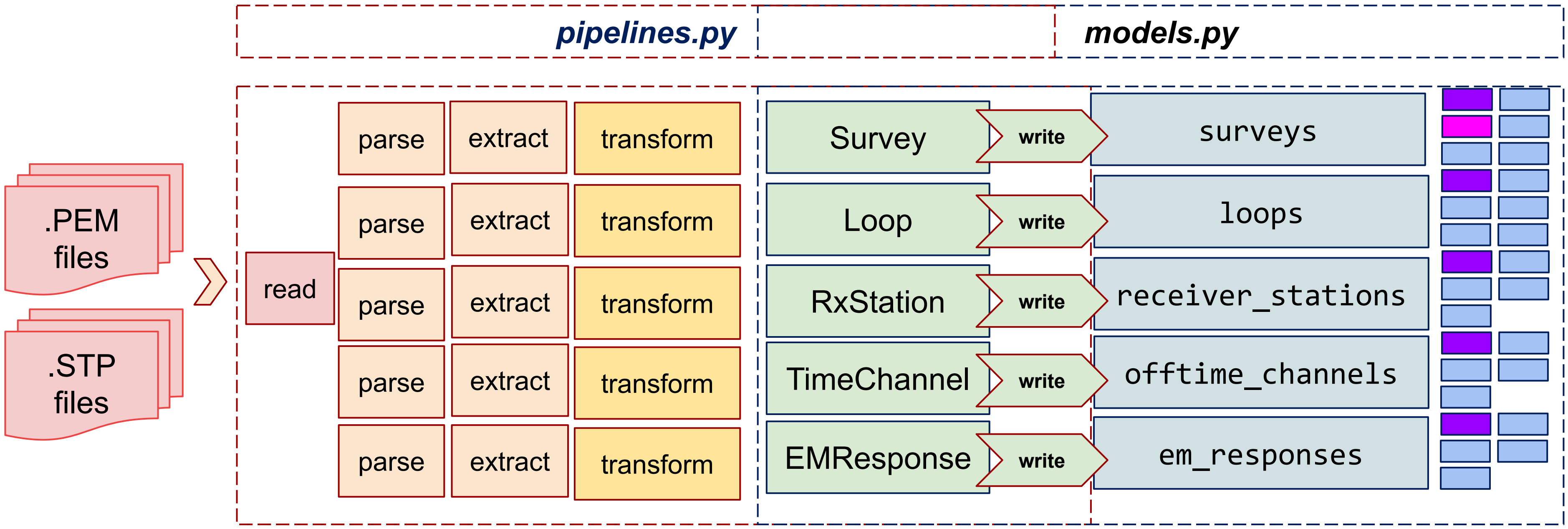


Basic Components of Data Pipeline

ENVS 5726

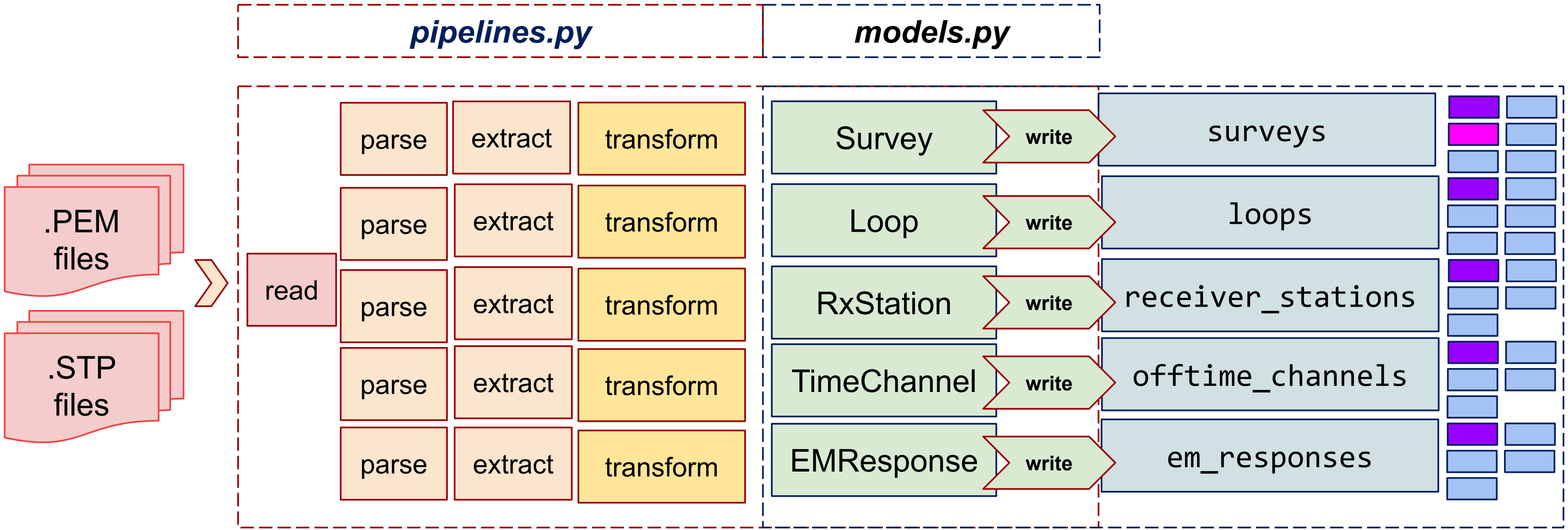


Basic Components of Data Pipeline



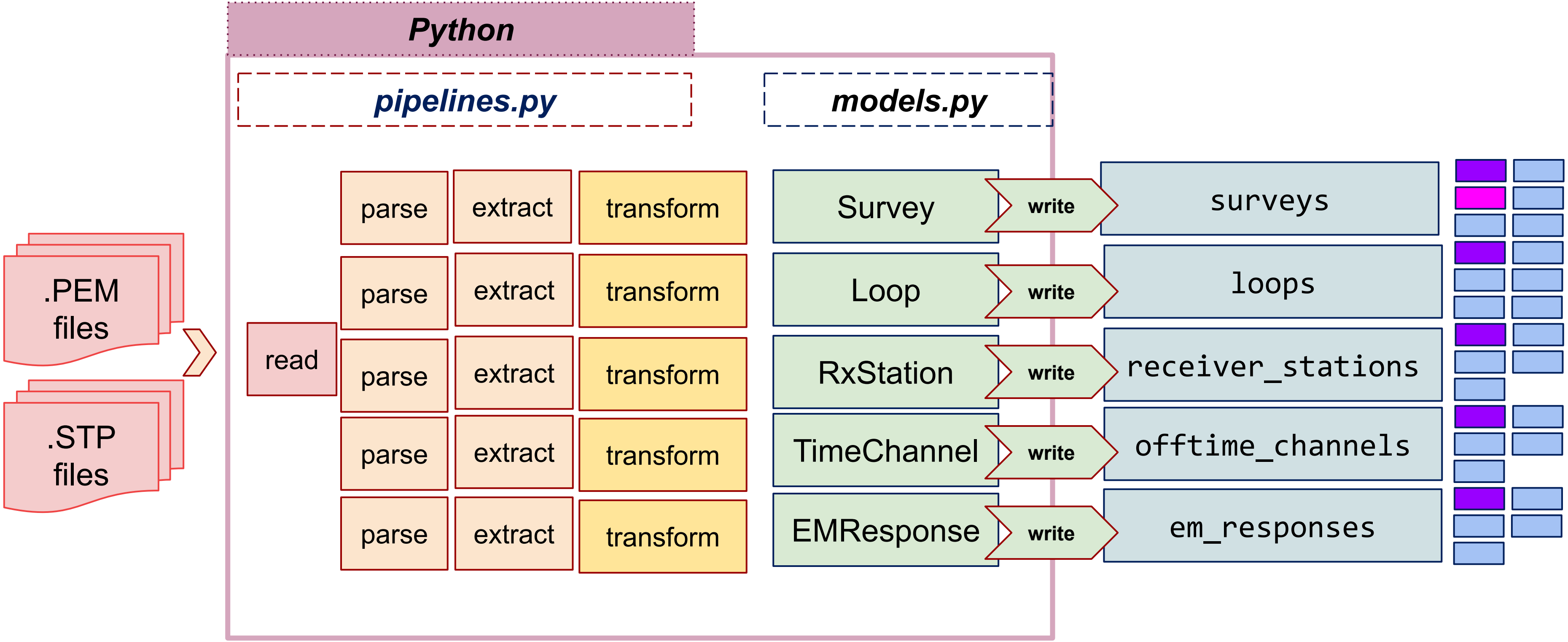
Basic Components of Data Pipeline

ENVS 5726

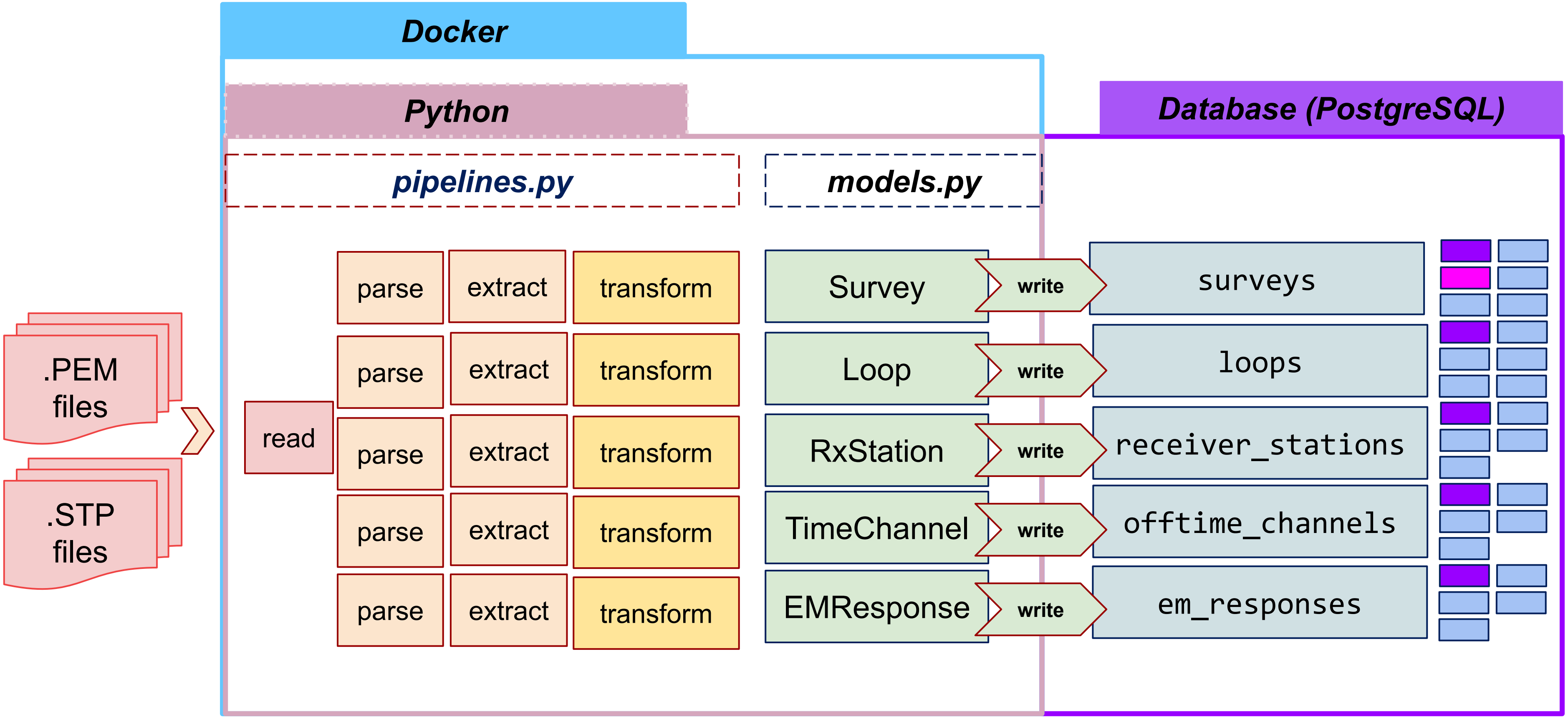


Basic Components of Data Pipeline

ENVS 5726

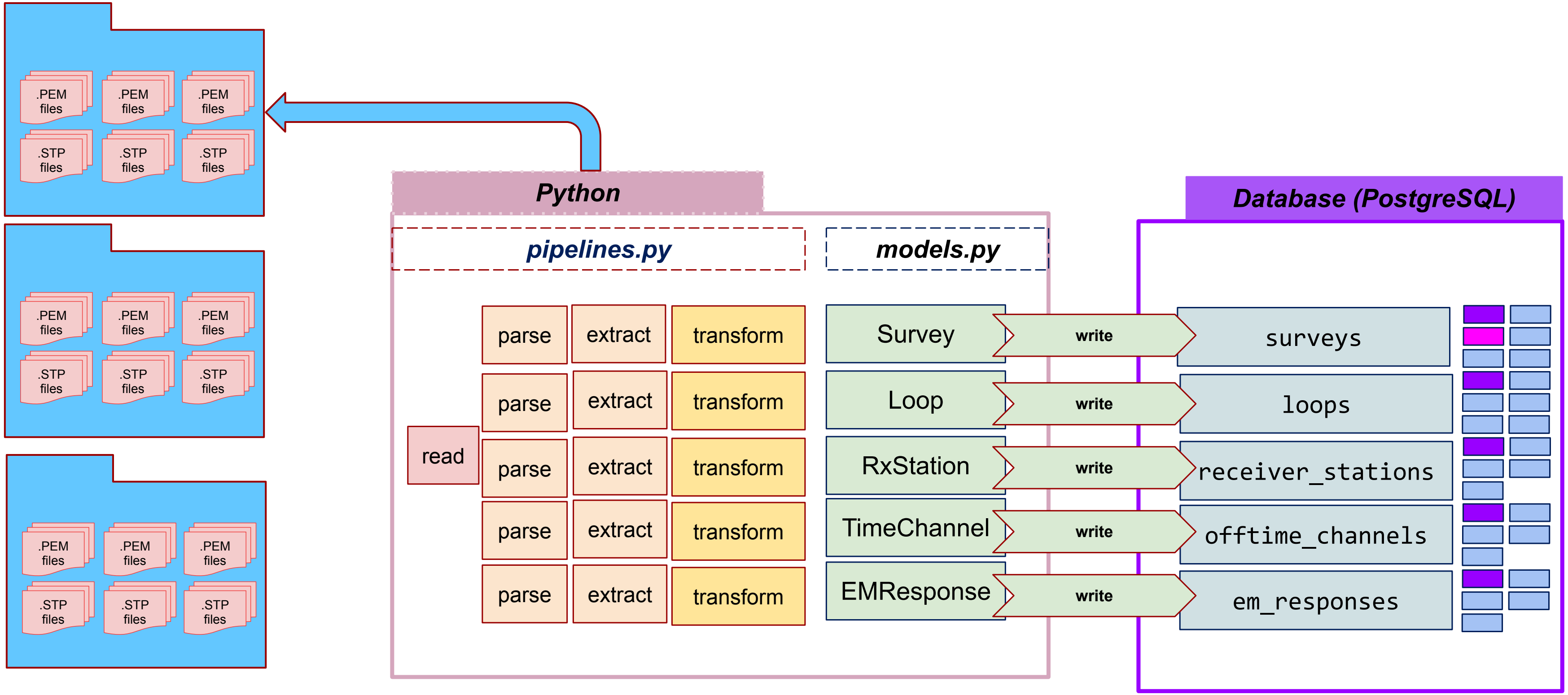


Basic Components of Data Pipeline



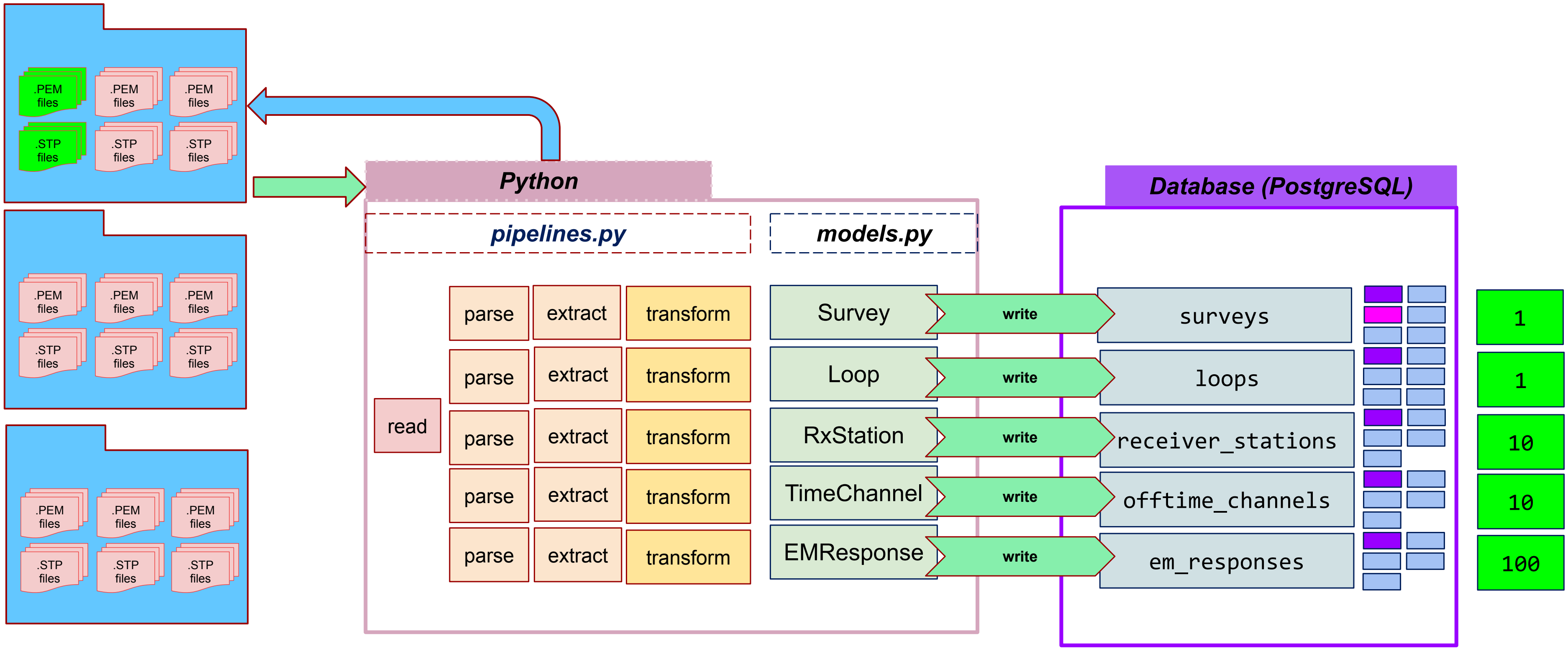
Scaling Data Pipeline

ENVS 5726

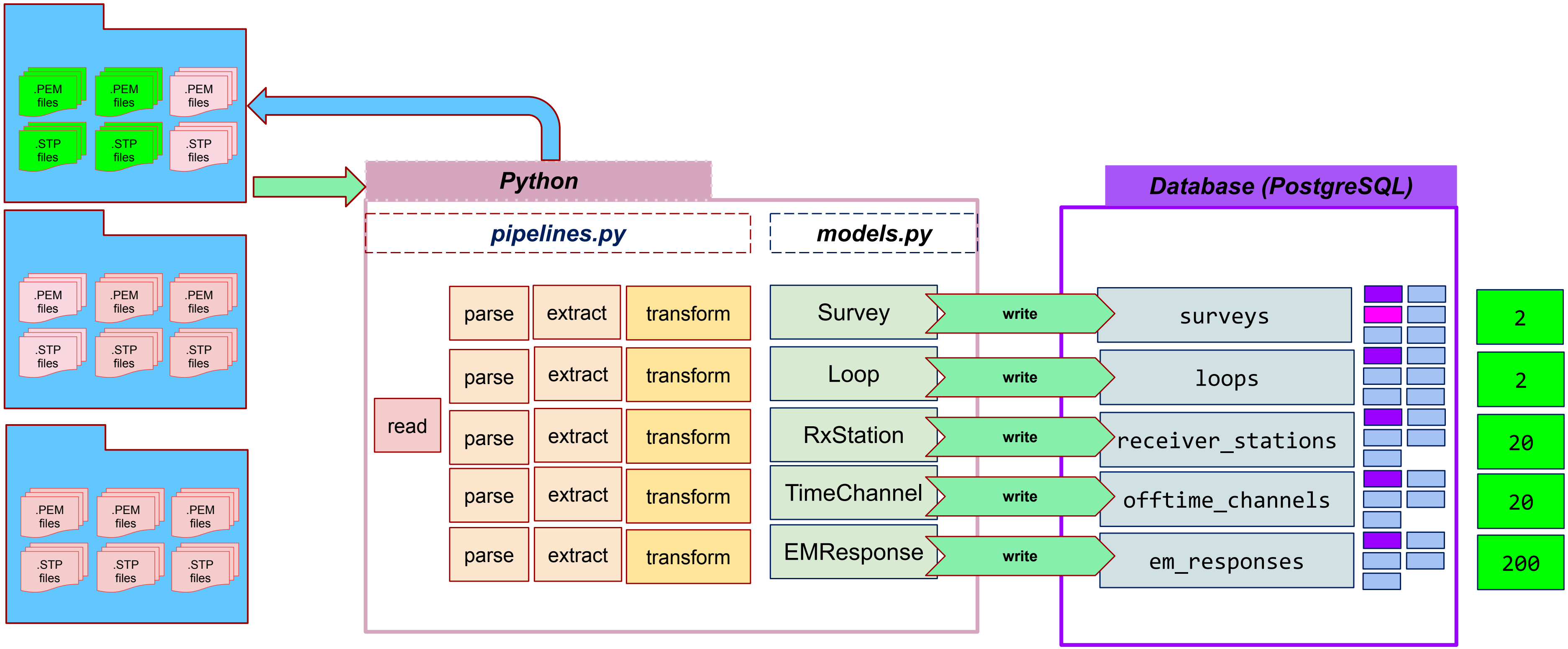


Scaling Data Pipeline

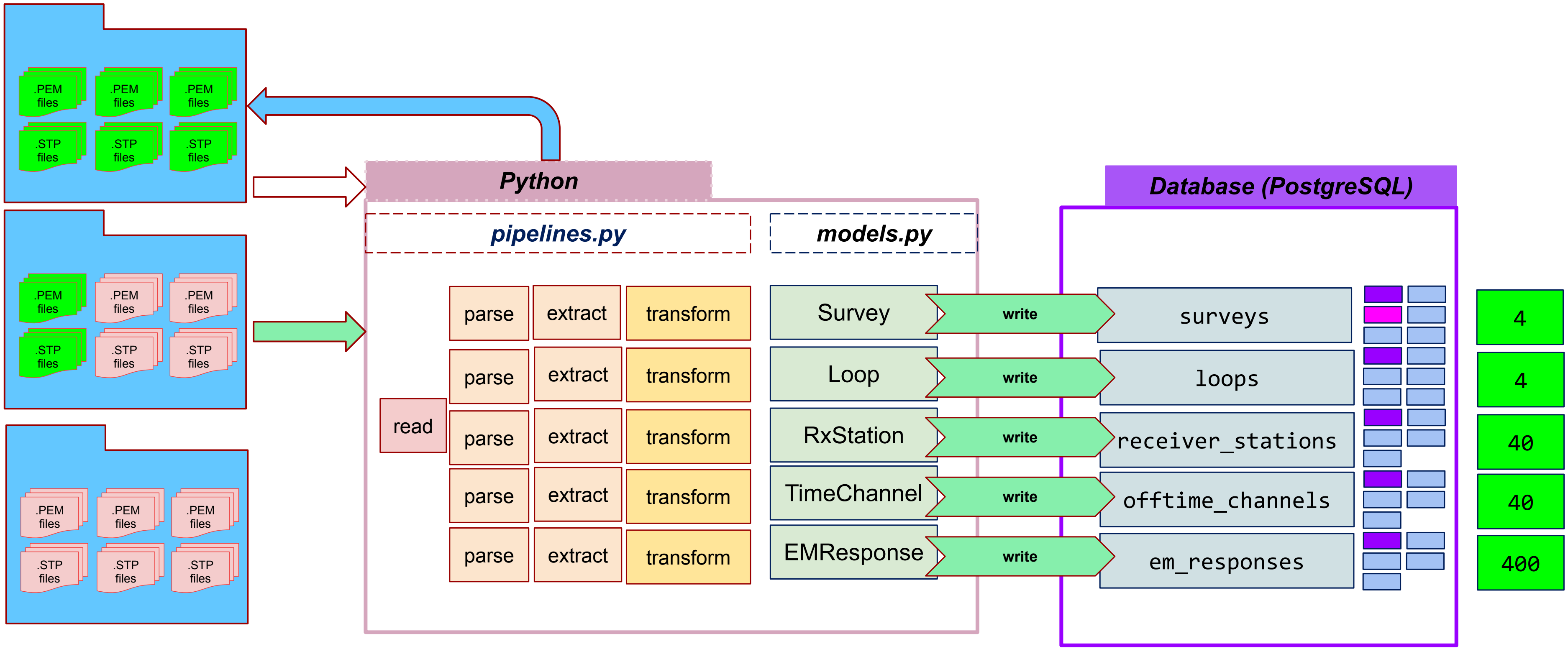
ENVS 5726



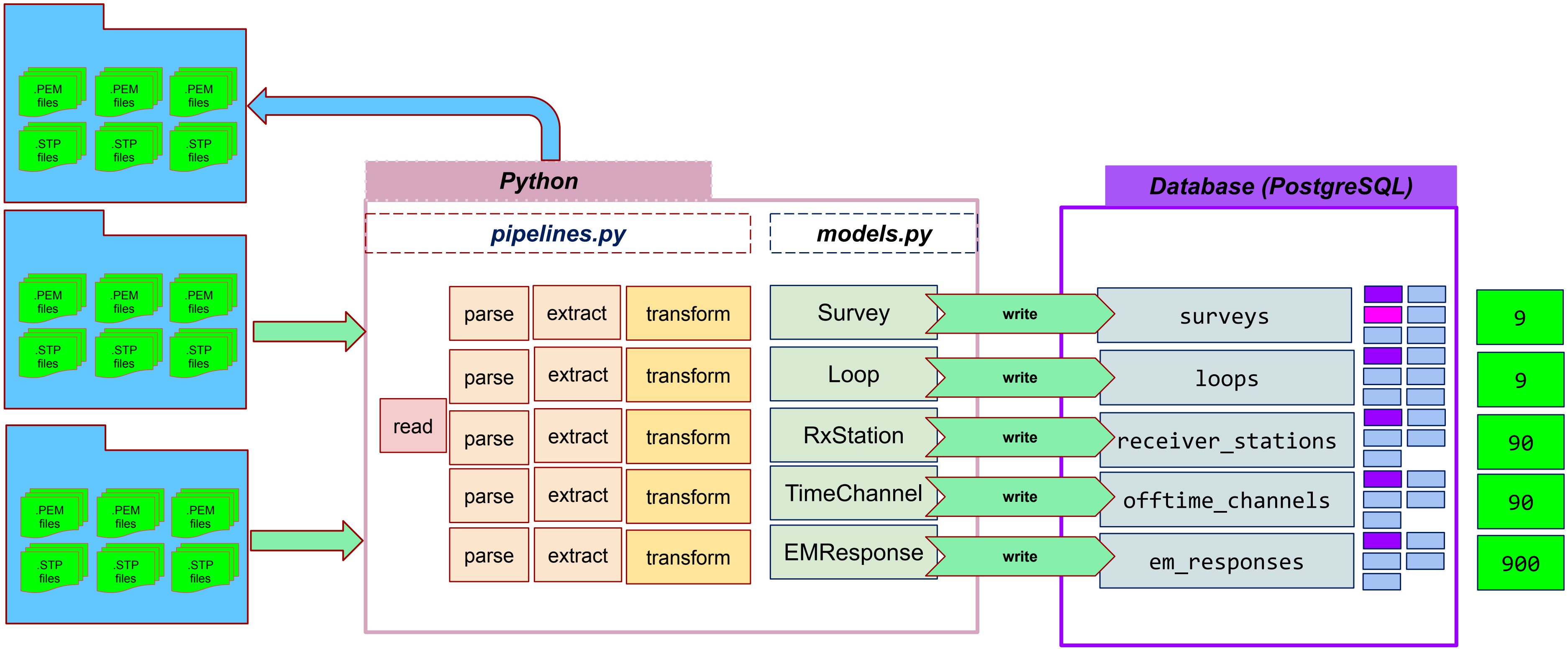
Scaling Data Pipeline



Scaling Data Pipeline



Scaling Data Pipeline

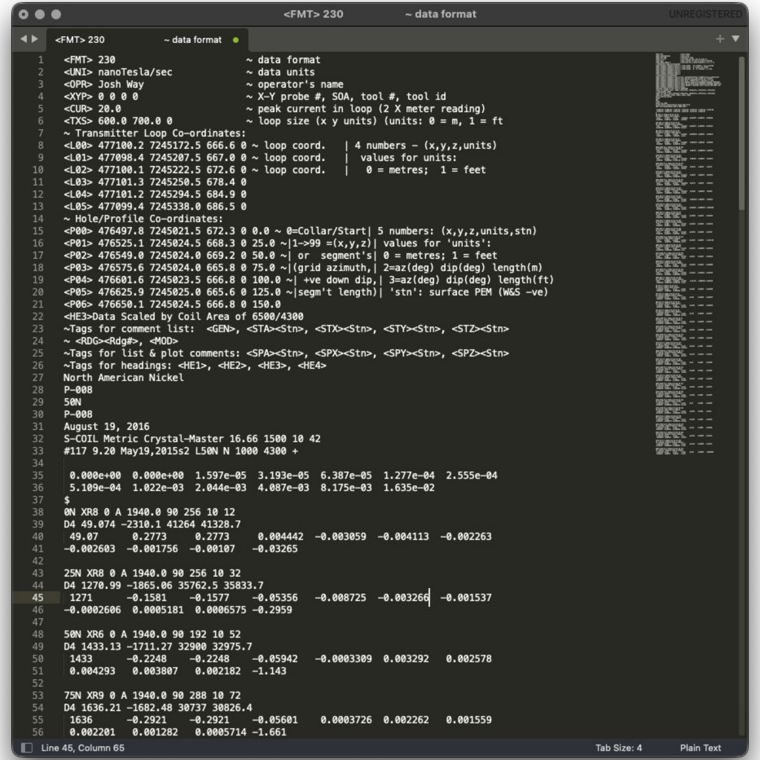


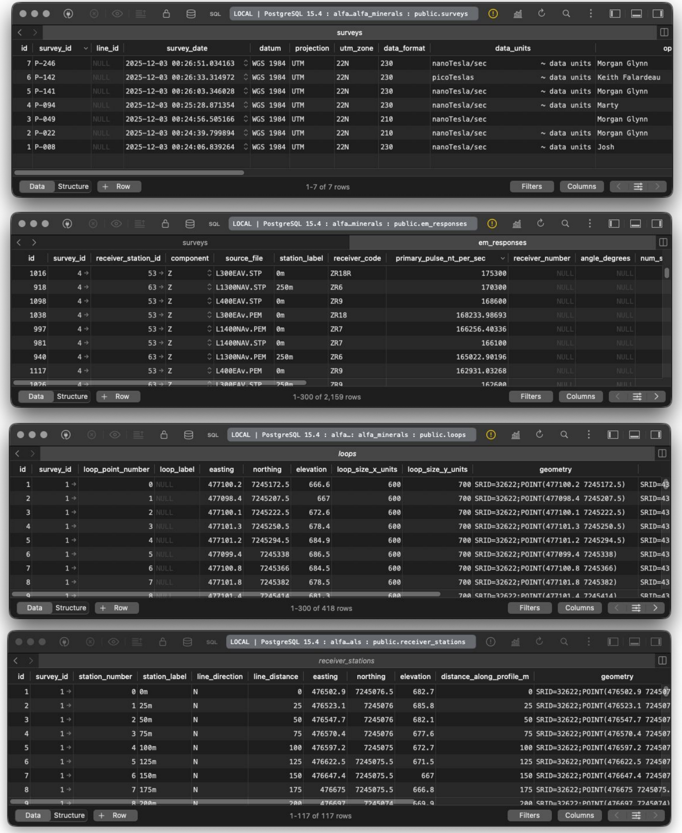
04

Demo Check-In

Demo time!

ENVS 5726





Data Engineering Pipelines

- **Scalable Task Design with Deduplication Checks:** Each save function (save_survey, save_loops, save_receiver_stations) includes upfront checks that query for existing records before inserting. This prevents duplicate data ingestion if the pipeline re-runs, returning early with counts of existing records which is a critical pattern for reliable ETL pipelines.

- **Batch Processing with Scalable Looping Patterns:** The code processes large datasets efficiently by iterating through lists of parsed records (loops, stations, measurements) and committing them in controlled batches. Rather than individual commits per record, functions like save_receiver_stations loop through all stations, add them to the session, and commit once thus reducing database round-trips and enabling the pipeline to scale to thousands of measurements per file.

- **Rich Data Extraction with Structured Parsing:** The parsing tasks (e.g., parse_measurements, parse_stations) use regex patterns and line-by-line iteration to extract complex, nested data structures from unstructured file content. Each function returns well-typed dictionaries with metadata context (line names, component types), ensuring downstream storage functions receive clean, validated data ready for database insertion.

Data Engineering Models and Schema

- **Robust Data Integrity with Cascades:** The models explicitly define parent-child relationships using *cascade="all, delete-orphan"*. This ensures that when a parent record (like a survey) is deleted, all associated children (loops, stations, measurements) are automatically cleaned up, preventing orphaned records in the database.

- **Explicit Database Constraints and Indexing:** proactively enforces data quality and performance through __table_args__. It defines UniqueConstraint to prevent logical duplicates (e.g., ensuring a specific loop_point_number only appears once per survey) and creates specific Index entries (e.g., on easting and northing) to speed up lookups.

- **Intentional Typing and Spatial Awareness:** The models go beyond basic types by integrating specialized fields like Enum for fixed categories (e.g., ComponentEnum) and PostGIS Geometry columns with specific SRIDs (Spatial Reference System Identifiers). This ensures the database enforces valid values and correctly handles complex spatial data rather than relying solely on application logic.

05

Questions and Conclusions

Conclusion

Knowledge about underlying data and user objective is critical to design a resilient and effective data pipeline.

A thoughtful schema design can go a long way (quality, resilience, and reliability).

Off-the-shelf orchestration tools are critical for pipeline observability.

Learning and applying the fundamentals (earth science, data engineering, software development) is critical despite LLMs.

Get feedback from users and stakeholders, and incorporate that feedback into the logic.

Discussion

Questions

fin