

ENVS 5726 Guest Lecture

Building Data Pipelines to Support Mineral Exploration Activity

Abdel Alfahham
alfahham@sas.upenn.edu

Founding Data Engineer

TABLE OF CONTENTS

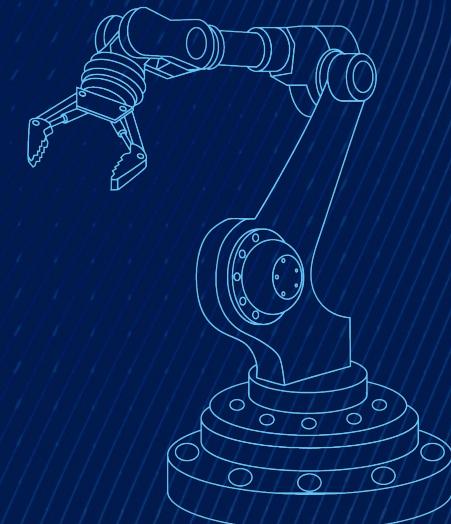
01 Introduction

02 Geophysical Surveys Fundamentals

03 Building Data Pipelines

04 Demo

05 Conclusion



01

Introduction

Summary and Objective

Career Timeline

ENVS 5726

- BS Environmental Sciences
- Environmental Consulting
- Penn MSAG
- Data Engineer

[2009 - 2013]
[2013 - 2018]
[2016 - 2019]
[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

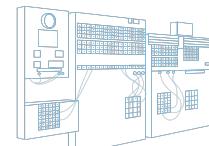
 **KoBold Metals**
Full-time · 2 yrs 11 mos
San Francisco Bay Area · Remote

 **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.

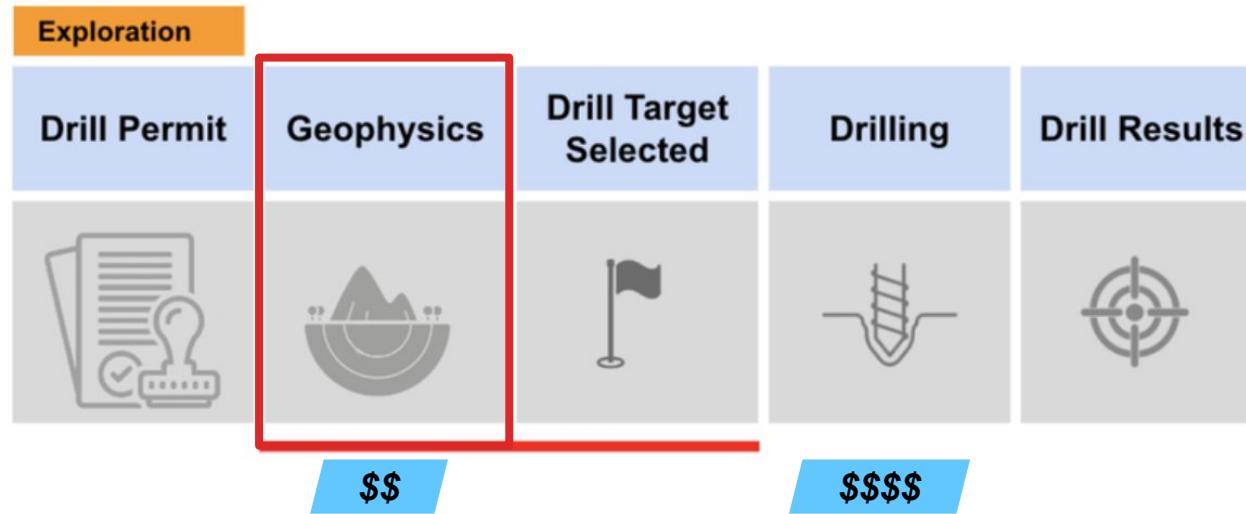
02

Geophysical Surveys

Ground EM Fundamentals

Geophysics in Mineral Exploration

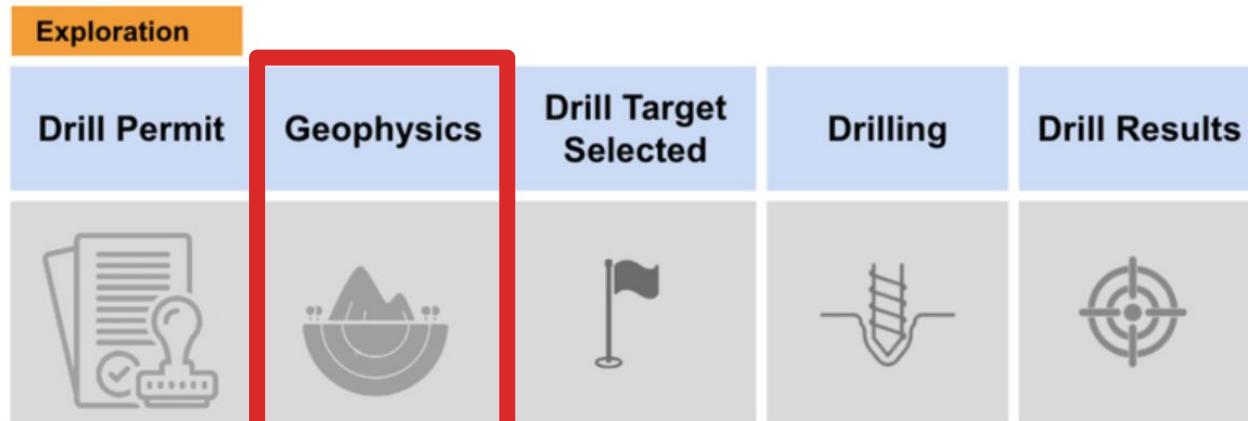
ENVS 5726



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

ENVS 5726



**Electromagnetics
(EM)**

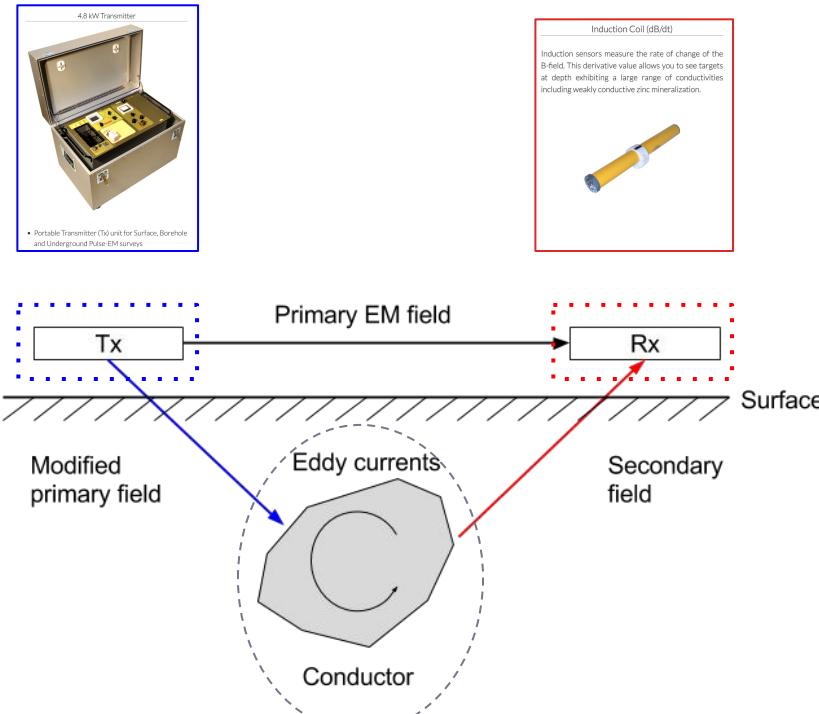
Gravity

**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

ENVS 5726

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)

ENVS 5726

- **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:** If 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

ENVS 5726

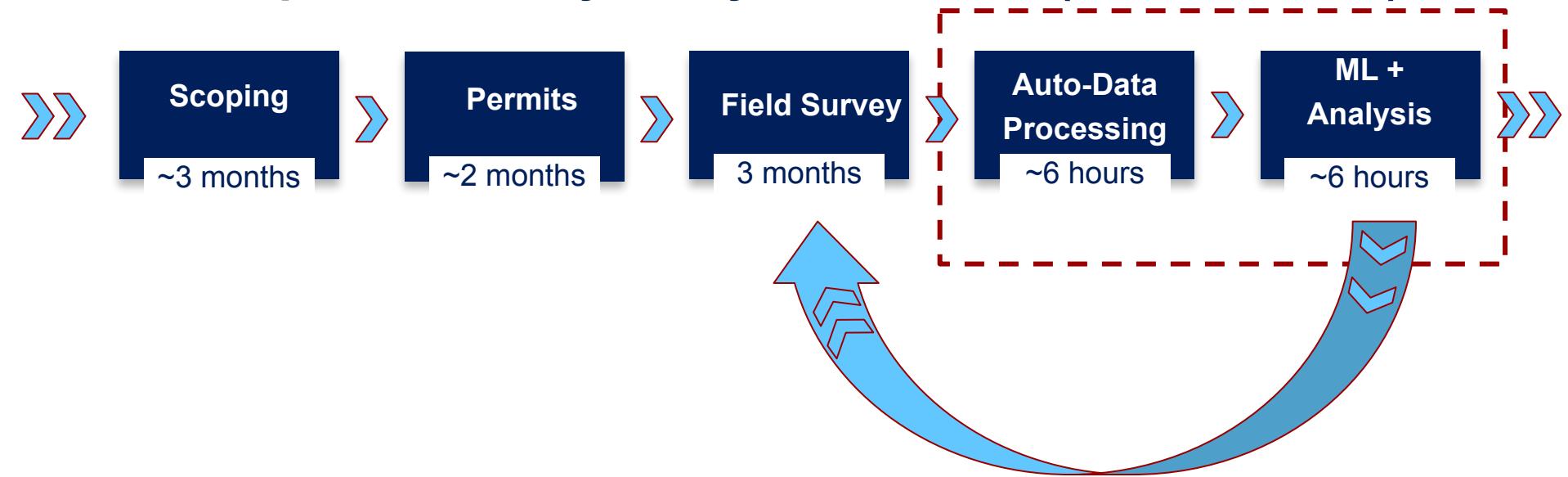
Conventional Survey Analysis Pace (12 Months)



Electromagnetic (EM) Survey Timeline

ENVS 5726

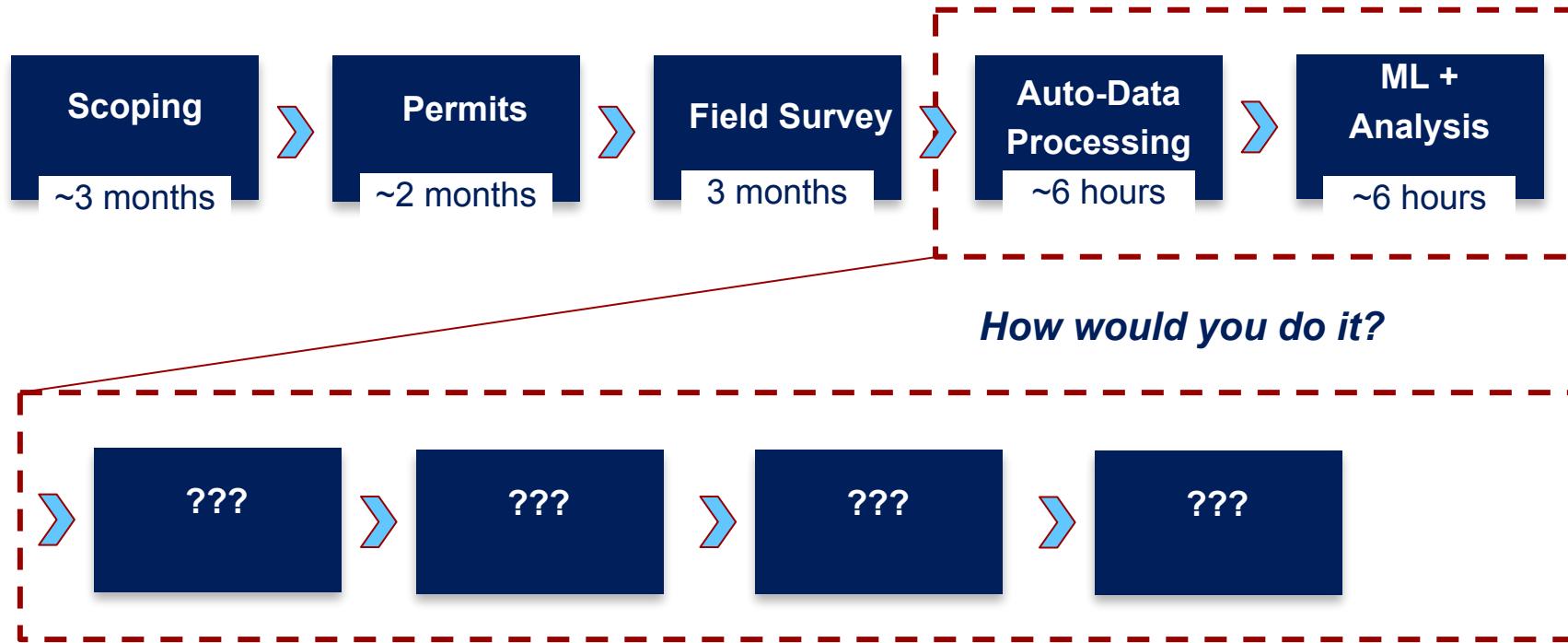
Proposed Survey Analysis Timeline (Fast Iteration)



Electromagnetic (EM) Survey Timeline

ENVS 5726

Proposed Survey Analysis Timeline (Fast Iteration)



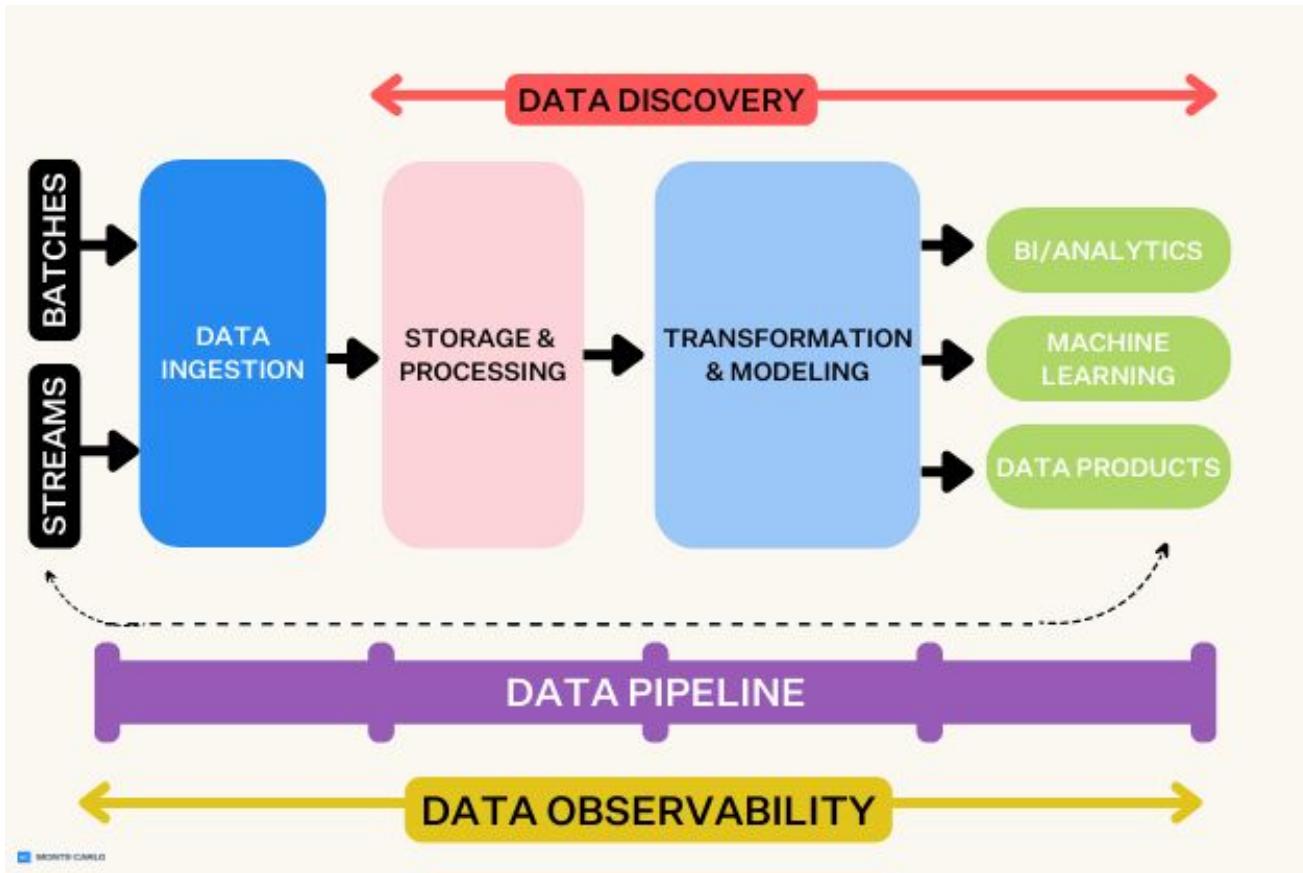
03

Data Pipelines

Automation, Scaling, Design

Basic Components of Data Platform

ENVS 5726



Let's Start the Demo Now!

ENVS 5726

```
<FMT> 230      - data format          UNREGISTERED
<FMT> 230      ~ data format
<INT> nanoTesla/sec
<OPR> Josh Way
<OPR> nanoTesla/sec
<OPR> operator's name
<XYP> 0 0 0
<CUR> 20.0
<TXS> 600.0 700.0 0
~ Transmitter Loop Co-ordinates
<L00> 477101.2 7245217.5 668.0 0 ~ loop coord. | 4 numbers -(x,y,z,units)
<L01> 477101.2 7245217.5 670.0 0 ~ loop coord. | values for units:
<L02> 477100.1 7245222.5 672.6 0 ~ loop coord. | 0 = metres; 1 = feet
<L03> 477101.3 7245259.5 678.4 0
<L04> 477101.2 7245294.5 684.9 0
<L05> 477099.4 7245338.0 686.5 0
~ Hole/Profile Co-ordinates:
<P00> 476497.8 7245021.6 672.3 0 ~ 0=Collar/Start| 5 numbers: (x,y,z,units,stn)
<P01> 476525.7 7245024.5 668.0 0 ~ 1=>99=(x,y,z,units) values for units:
<P02> 476525.7 7245024.5 668.0 0 ~ 2=segment length| ft
<P03> 476576.6 7245024.6 665.8 0 75.0 ~|(grid|length| 2az(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 7245020.6 665.6 0 125.0 ~|segmt length| stn: surface PEM (MGS -ve)
<P06> 476650.1 7245024.5 666.8 0 150.0
<H3>Data Scaled by Coil Area of 650/4300
<Tags for comment list: <GEN>, <TA><Stn>, <TX><Stn>, <TY><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
5N
P-008
August 19, 2016
S-COIL Metric Crystal-Master 16.66 1500 10 42
#17 9.20 May19,2015z2 150W N 1000 4300 +
0.000e+00 0.000e+00 1.597e-05 3.193e-05 6.387e-05 1.277e-04 2.555e-04
5.189e-04 1.022e-03 2.044e-03 4.087e-03 8.175e-03 1.635e-02
$ 
ON XRB Ø A 1940.0 98 256 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 XRB Ø A 1940 98 256 10 32
D4 1270.99 -1865.06 35762.5 35833.7
1271 -0.1581 -0.1577 -0.05356 -0.008725 -0.003266 -0.001537
-0.000266 0.0005181 0.0006575 -0.2959
5N XRB Ø A 1940.0 98 192 10 52
D4 1433.13 -1711.27 3298 3297.7
1433 -0.2248 -0.2248 -0.05942 -0.0003389 0.003292 0.002578
0.004293 0.008387 0.002182 -1.143
75N XRB Ø 25 98 288 10 72
D4 1636.21 -1682.48 38737 38826.4
1636 -0.2921 -0.2921 -0.05601 0.0003726 0.002262 0.001559
0.002201 0.001282 0.0005714 -1.661
Line 45, Column 65
```

Tab Size: 4 Plain Text



surveys									
id	survey_id	line_id	survey_date	datum	projection	utm_zone	data_format	data_units	op
7	P-246		2825-12-03 00:26:51.834163	MGS	1984	UTM	22N	238	nanoTesla/sec
6	P-142		2825-12-03 00:26:33.319497	MGS	1984	UTM	22N	238	picoTesla
5	P-141		2825-12-03 00:26:03.346928	MGS	1984	UTM	22N	238	nanoTesla/sec
4	P-094		2825-12-03 00:25:26.871354	MGS	1984	UTM	22N	238	nanoTesla/sec
3	P-849		2825-12-03 00:24:56.585166	MGS	1984	UTM	22N	218	nanoTesla/sec
2	P-822		2825-12-03 00:24:49.799994	MGS	1984	UTM	22N	218	nanoTesla/sec
1	P-886		2025-12-03 00:24:46.639354	MGS	1984	UTM	22N	238	nanoTesla/sec

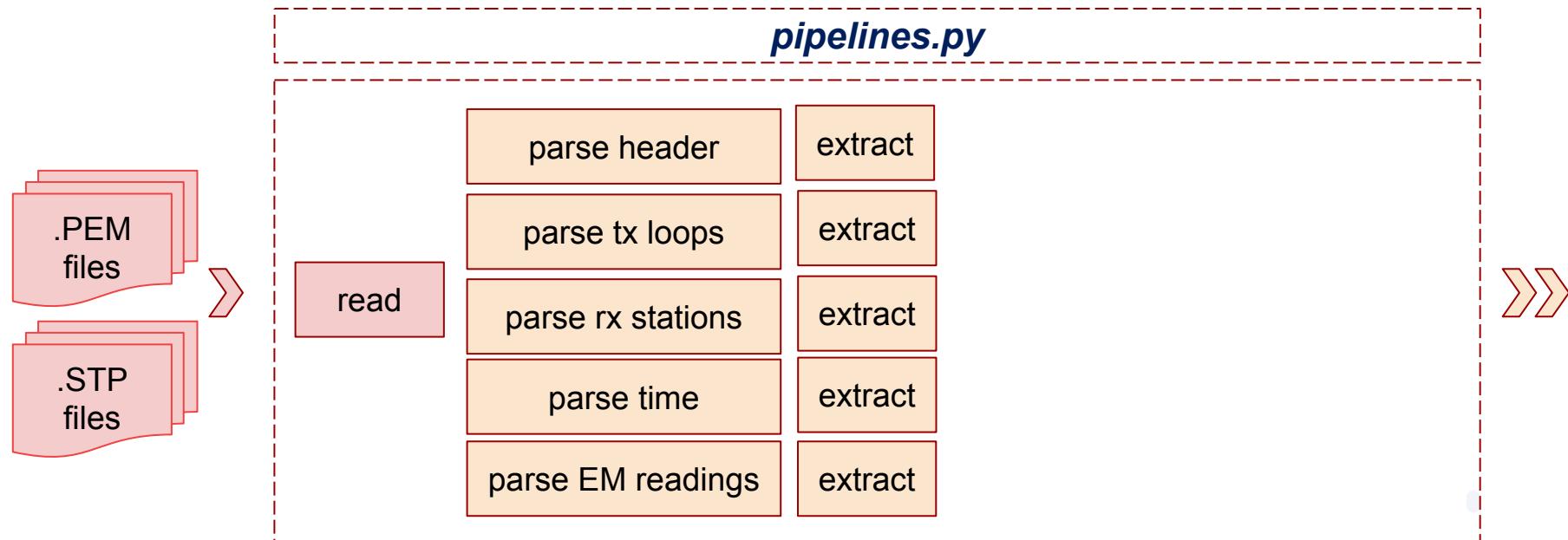
em_responses										
id	survey_id	receiver_station_id	component	source_file	station_label	receiver_code	primary_pulse_nt_per_sec	receiver_number	angle_degrees	num_c
1016	4	53	Z	L308EAV.STP	On	ZH1R	175300	NULL	NULL	NULL
918	4	63	Z	L308EAV.STP	25m	ZR6	170300	NULL	NULL	NULL
1098	4	53	Z	L408EAV.STP	On	ZR9	168600	NULL	NULL	NULL
1638	4	53	Z	L308EAV.PDF	On	ZB18	168233.99693	NULL	NULL	NULL
997	4	53	Z	L1408EAV.PDF	On	ZR7	166250.40335	NULL	NULL	NULL
981	4	53	Z	L1408EAV.STP	On	ZR7	165100	NULL	NULL	NULL
940	4	63	Z	L308EAV.PDF	25m	ZR6	165022.90165	NULL	NULL	NULL
1117	4	53	Z	L408EAV.PDF	On	ZR9	162931.83268	NULL	NULL	NULL
1099	4	53	Z	L308EAV.PDF	25m	ZR9	162600	1K3640	1K3640	1-300 of 2,159 rows

loops									
id	survey_id	loop_point_number	loop_label	easting	northing	elevation	loop_size_x_units	loop_size_y_units	geometry
1	1	1	Ø	477100.2	7245172.5	666.6	680	780	SRID=32622;POINT(477100.2 7245172.5)
2	1	1	1	477998.4	7245287.5	667	680	780	SRID=32622;POINT(477998.4 7245287.5)
3	1	1	2	477100.1	7245222.5	672.6	680	780	SRID=32622;POINT(477100.1 7245222.5)
4	1	1	3	477101.3	7245295.5	678.4	680	780	SRID=32622;POINT(477101.3 7245295.5)
5	1	1	4	477101.2	7245294.5	684.9	680	780	SRID=32622;POINT(477101.2 7245294.5)
6	1	1	5	477999.4	7245338	686.5	680	780	SRID=32622;POINT(477999.4 7245338)
7	1	1	6	477100.8	7245366	684.5	680	780	SRID=32622;POINT(477100.8 7245366)
8	1	1	7	477101.1	7245382	678.5	680	780	SRID=32622;POINT(477101.1 7245382)
9	1	1	8	477101.4	7245414	681.1	680	780	SRID=32622;POINT(477101.4 7245414)

receiver_stations										
id	survey_id	station_number	station_label	line_direction	line_index	easting	northing	elevation	distance_along_line_m	geometry
1	1	0	N	8	476582.9	7245678.5	662.7	6	580	SRID=32622;POINT(476582.9 7245678.5)
2	1	1	25m	N	58	476531.1	7245678.5	665.8	6	25 SRID=32622;POINT(476531.1 7245678.5)
3	1	1	2.5m	N	58	476547.7	7245678	662.1	58	SRID=32622;POINT(476547.7 7245678)
4	1	1	3.75m	N	58	476578.4	7245678	677.6	75	SRID=32622;POINT(476578.4 7245678)
5	1	1	4.10m	N	108	476597.2	7245678	672.7	100	SRID=32622;POINT(476597.2 7245678)
6	1	1	5.125m	N	125	476621.5	7245675.5	671.5	125	SRID=32622;POINT(476621.5 7245675.5)
7	1	1	6.15m	N	158	476641.4	7245675.5	667	158	SRID=32622;POINT(476641.4 7245675.5)
8	1	1	7.175m	N	175	476675	7245675.5	666.8	175	SRID=32622;POINT(476675 7245675.5)
9	1	1	8.98m	N	198	476697	7245674	660.0	198	SRID=32622;POINT(476697 7245674)

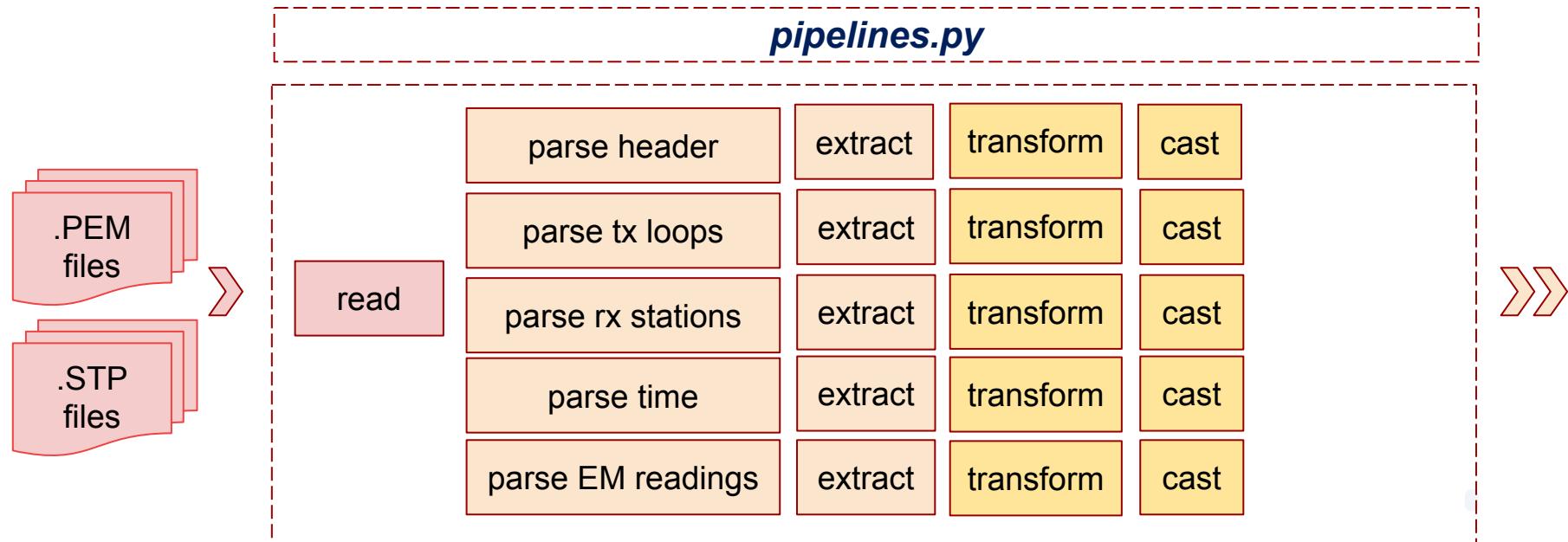
Basic Components of Data Pipeline

ENVS 5726



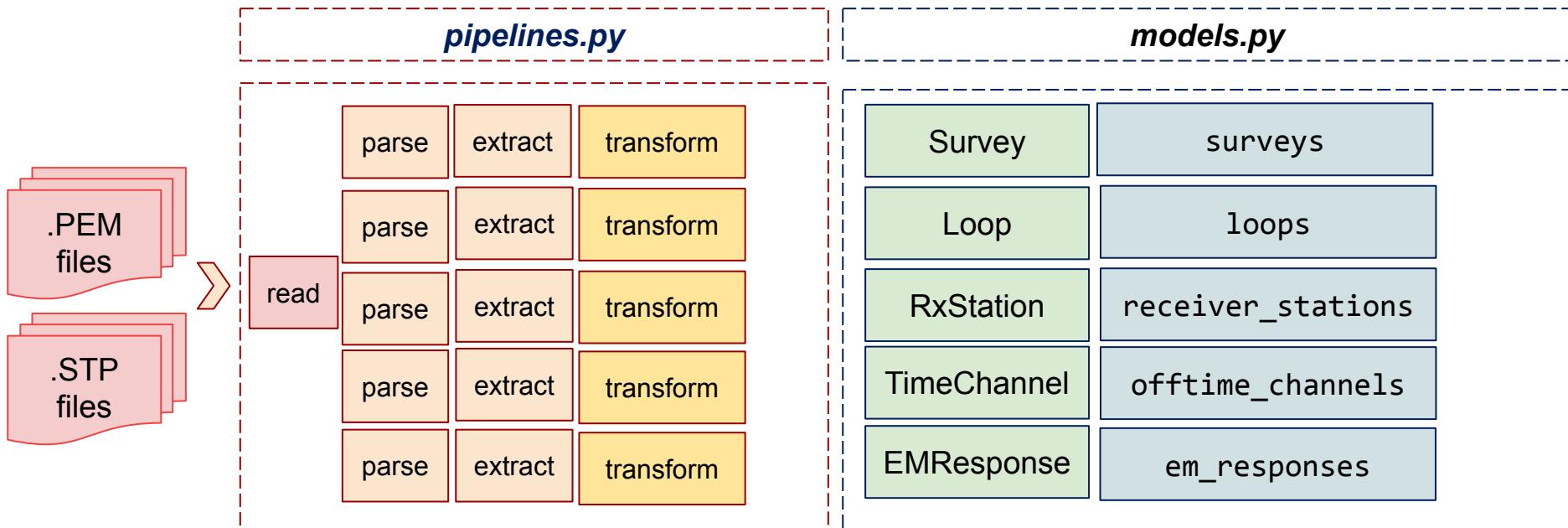
Basic Components of Data Pipeline

ENVS 5726



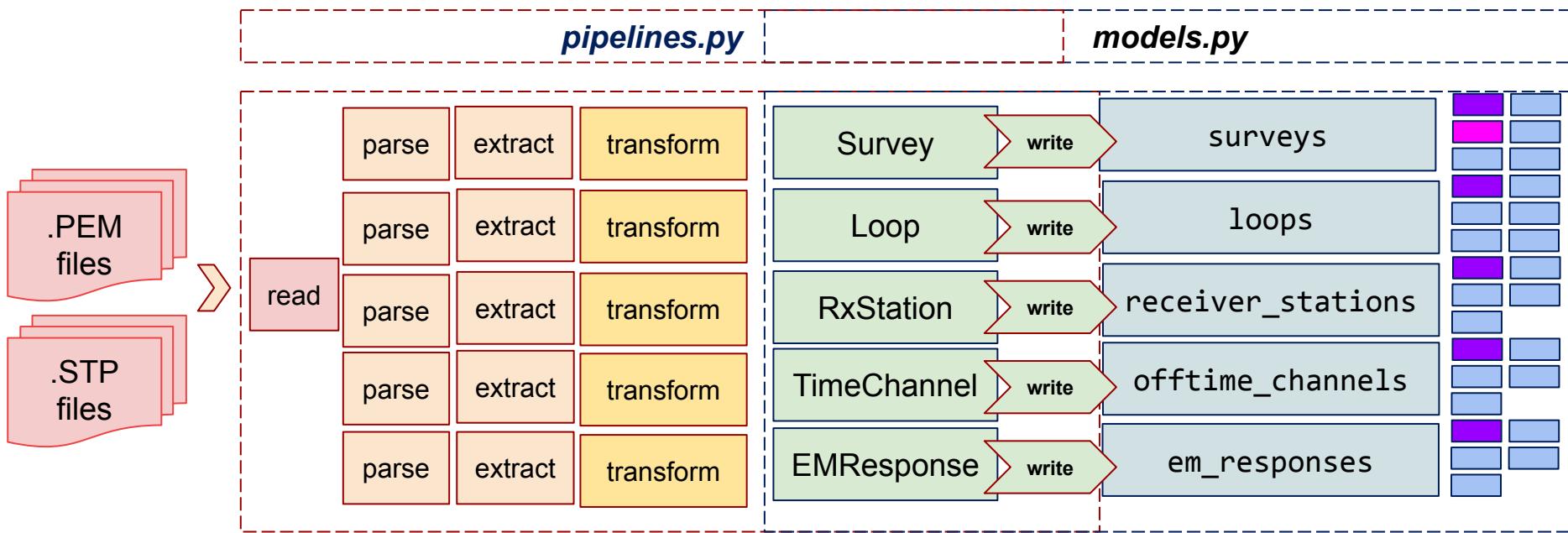
Basic Components of Data Pipeline

ENVS 5726



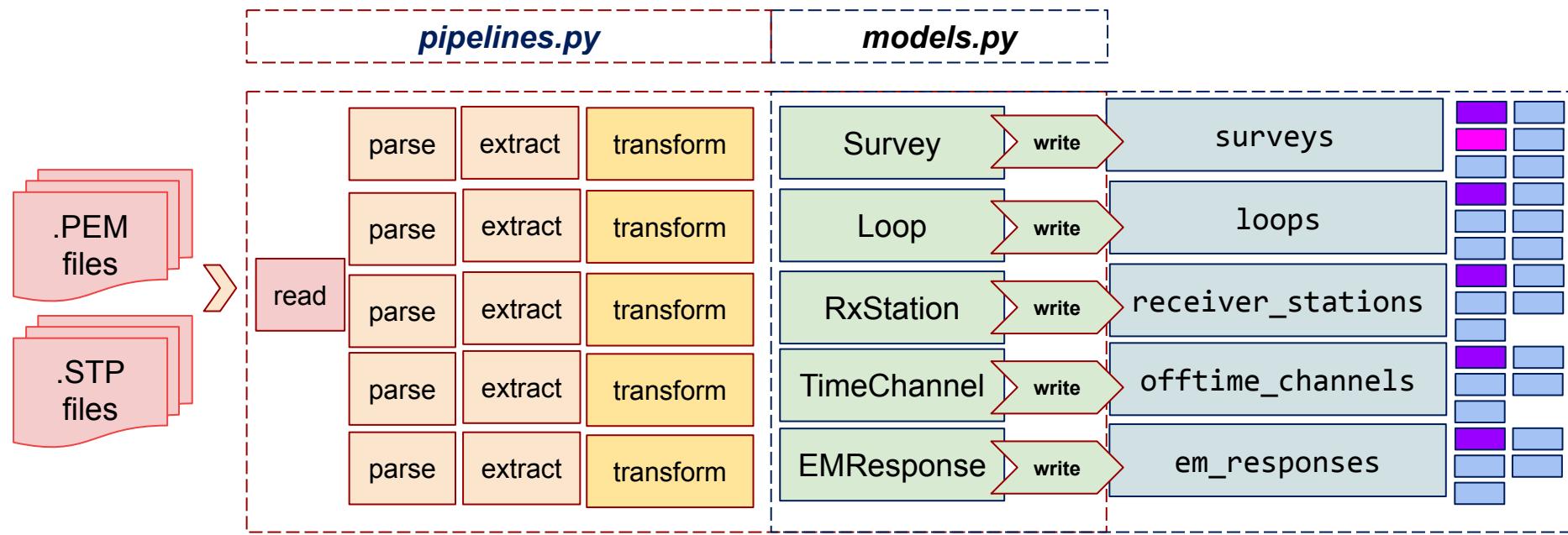
Basic Components of Data Pipeline

ENVS 5726



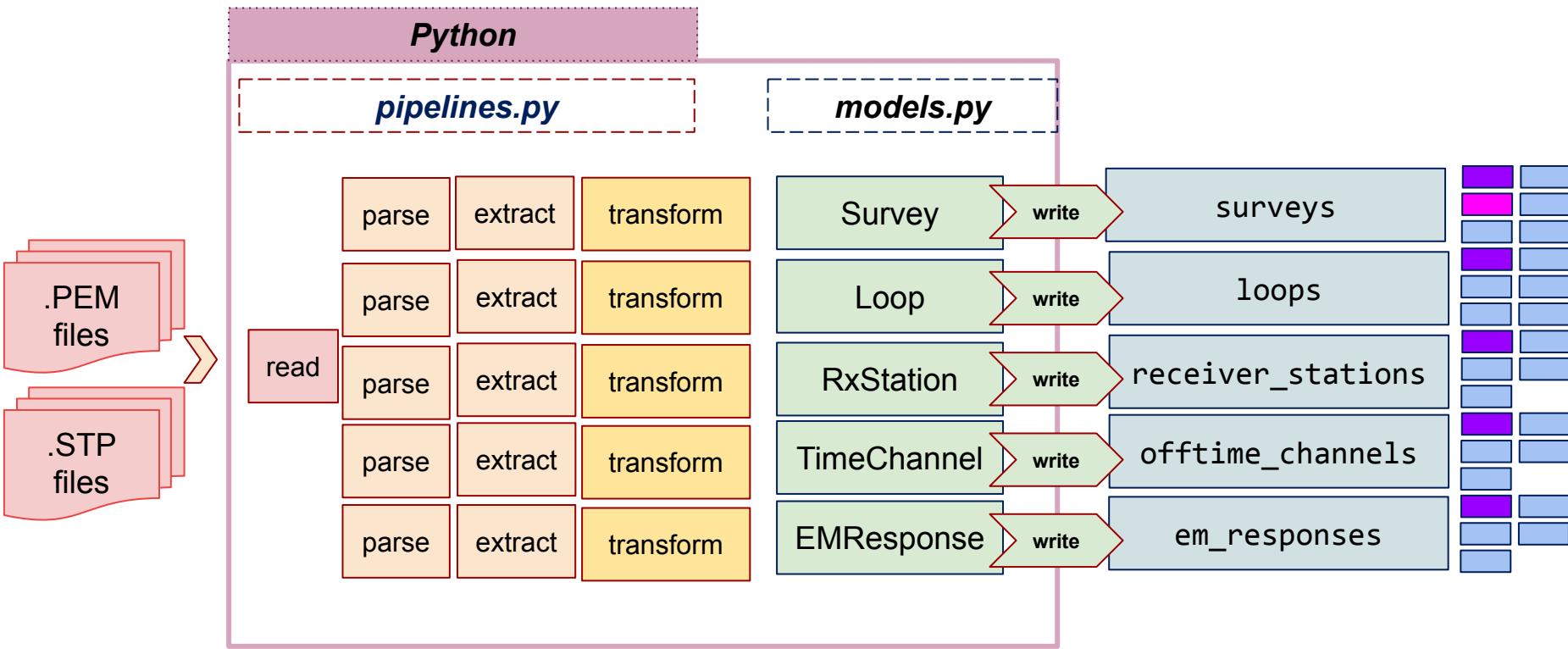
Basic Components of Data Pipeline

ENVS 5726



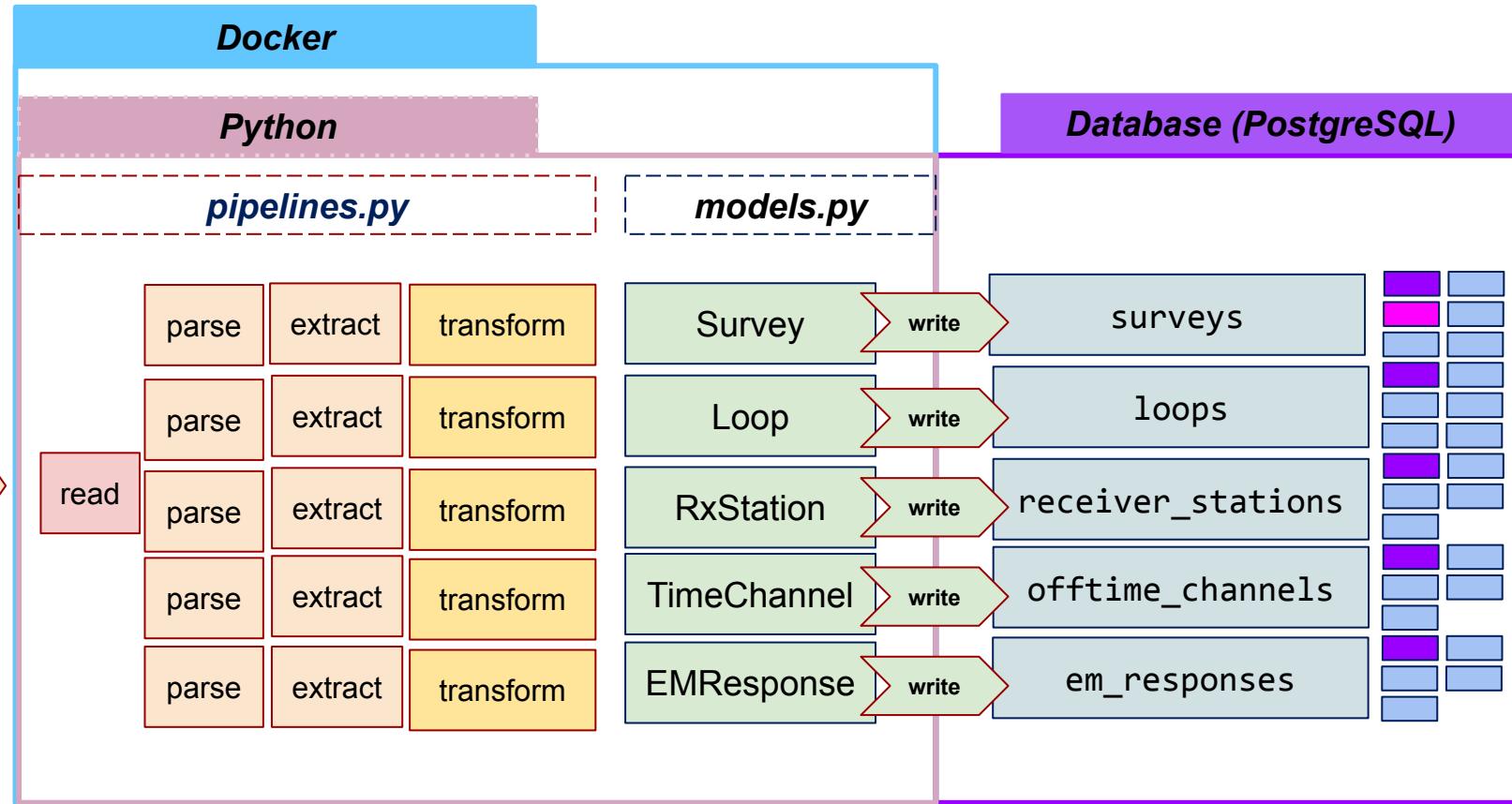
Basic Components of Data Pipeline

ENVS 5726



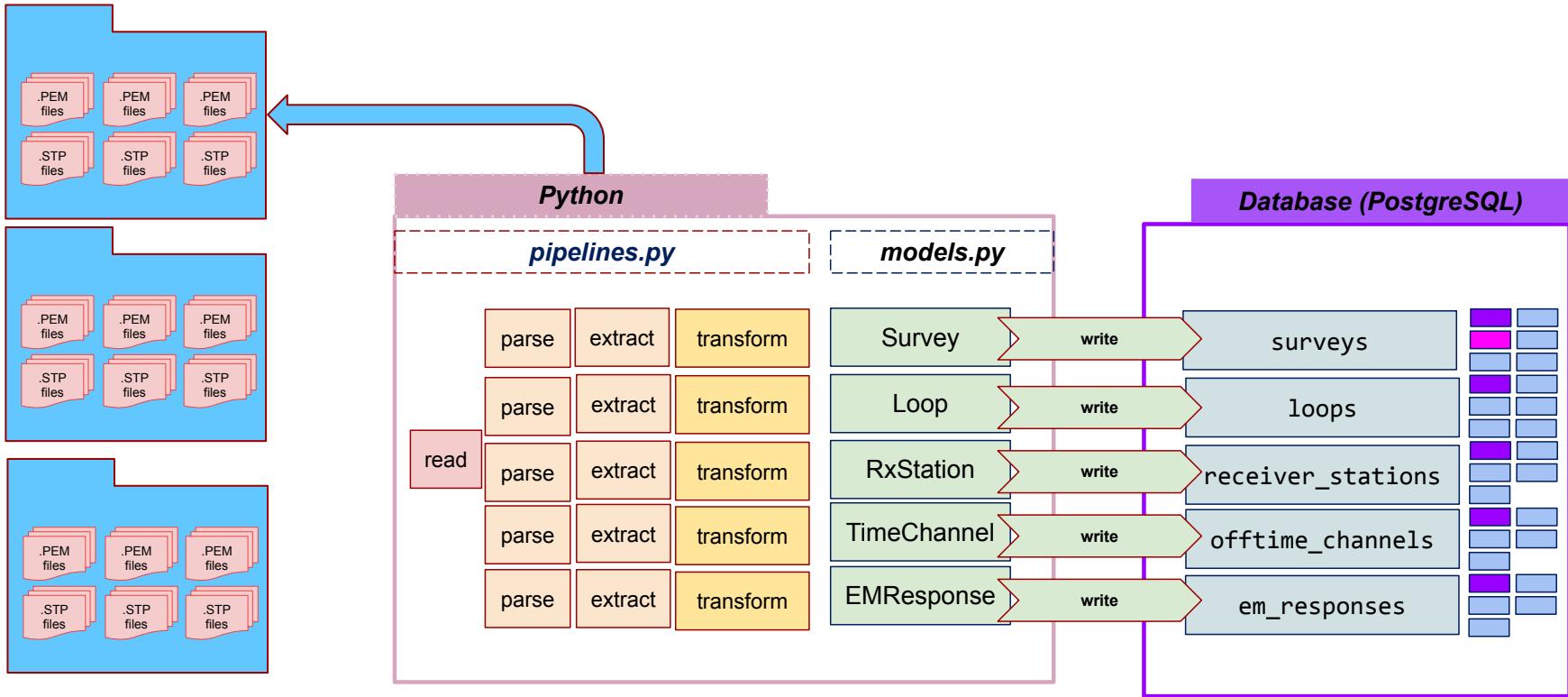
Basic Components of Data Pipeline

ENVS 5726



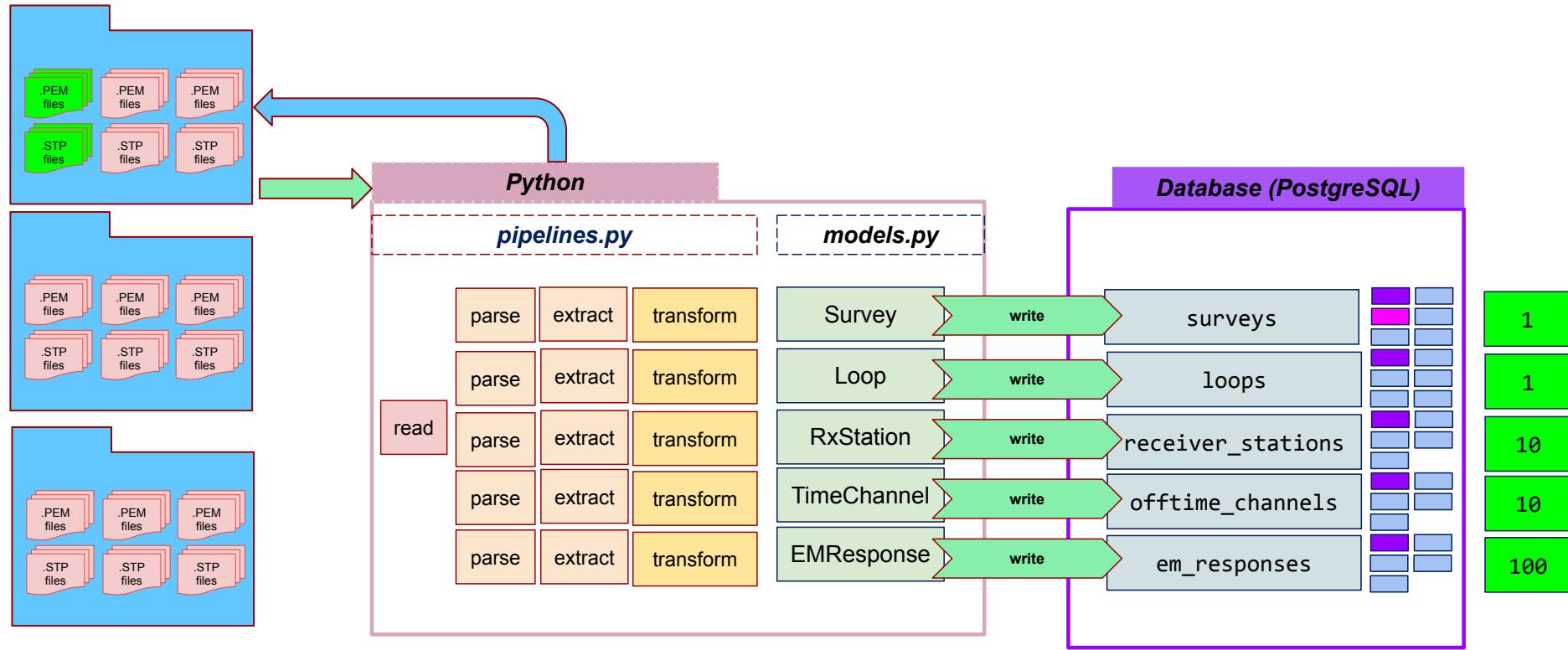
Scaling Data Pipeline

ENVS 5726



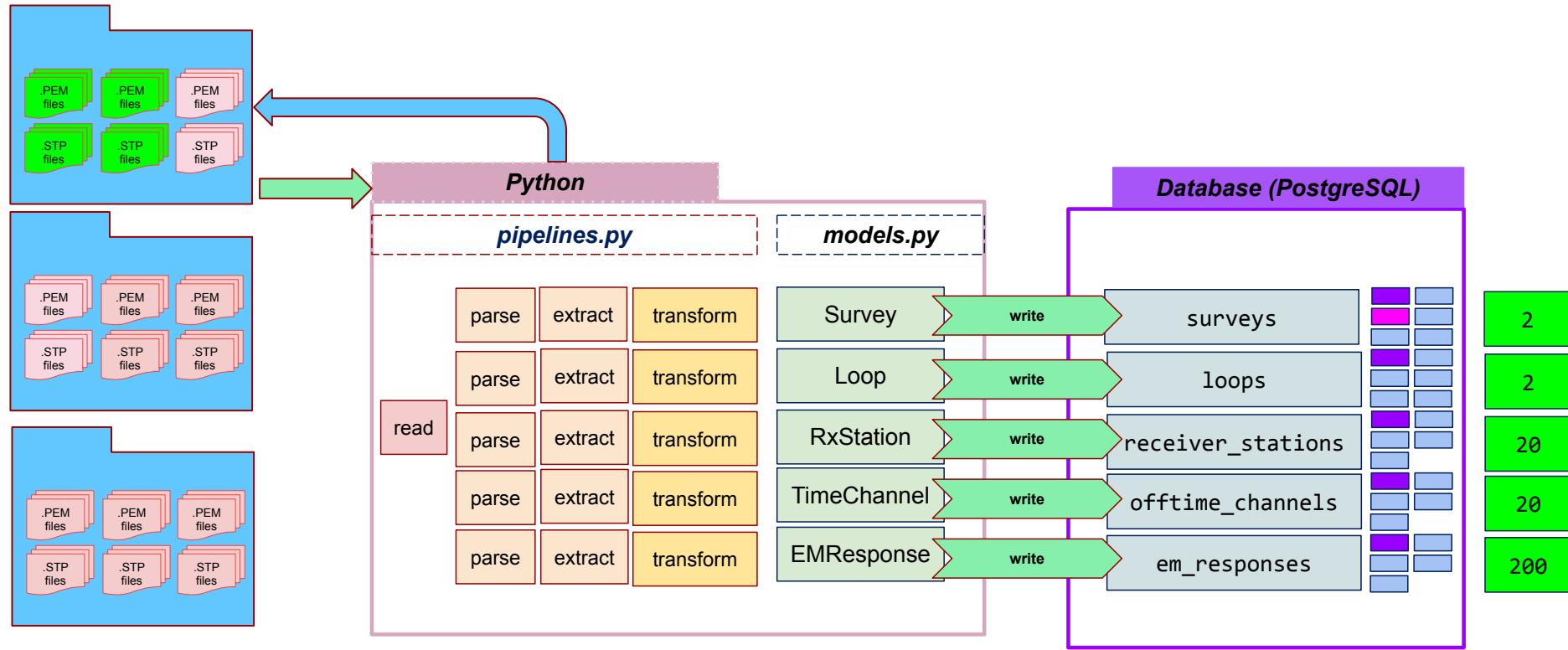
Scaling Data Pipeline

ENVS 5726



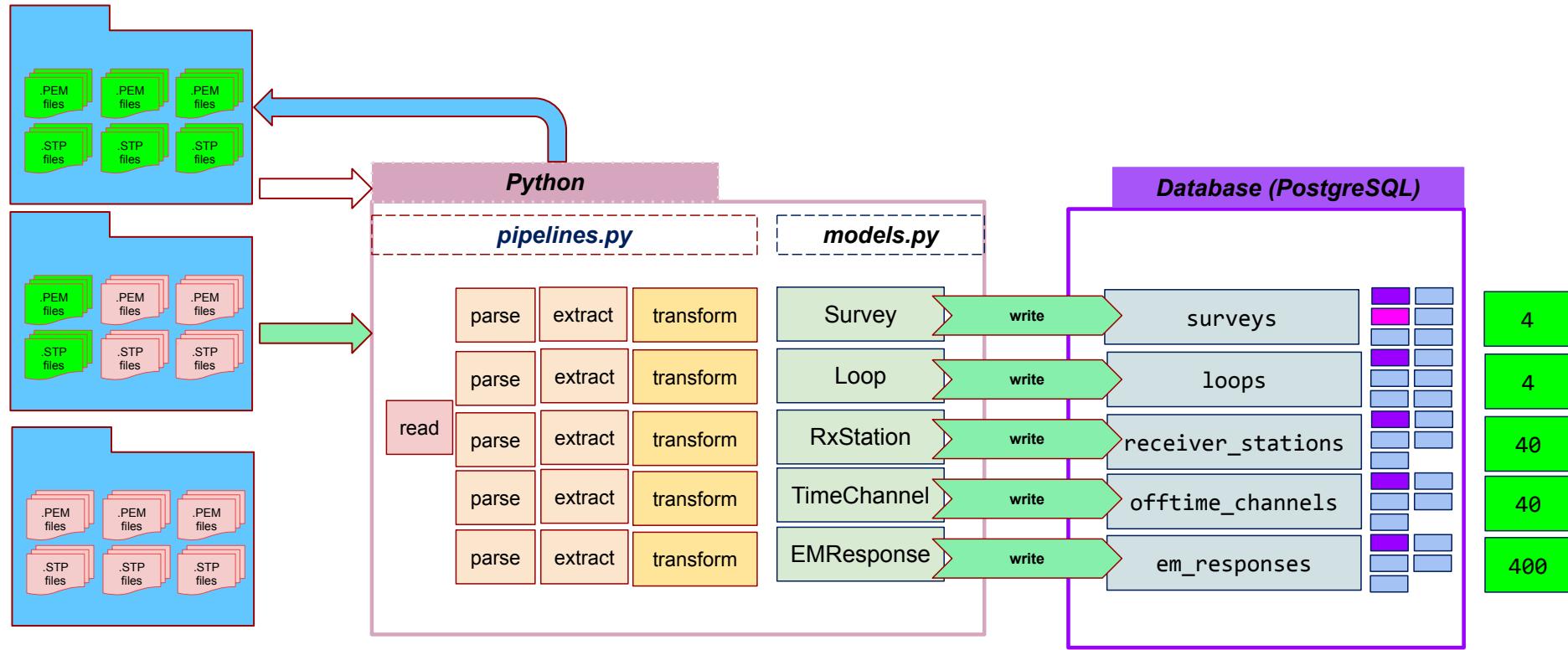
Scaling Data Pipeline

ENVS 5726



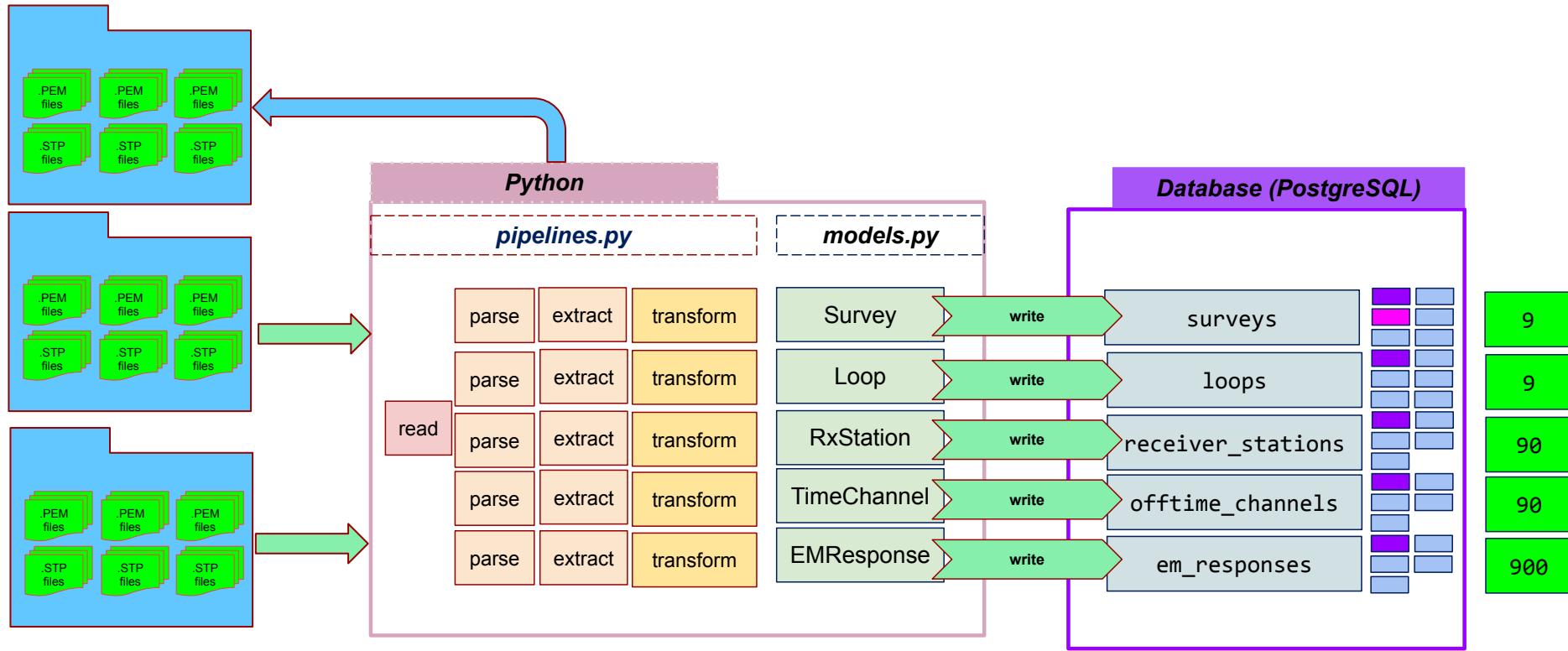
Scaling Data Pipeline

ENVS 5726



Scaling Data Pipeline

ENVS 5726



04

Demo Check-In

Demo time!

ENVS 5726

```
<FMT> 230      - data format          UNREGISTERED
1  <FMT> 230      ~ data format
2  <INT> nanoTesla/sec
3  <DPR> Josh Way
4  <XYP> 0 0 0
5  <CUR> 20.0
6  <TXS> 600.0 700.0 0
7  ~ Transmitter Loop Co-ordinates
8  <L0> 477101.2 7245217.5 600.0 ~ loop coord. | 4 numbers -(x,y,z,units)
9  <L0> 477101.2 7245217.5 672.0 ~ loop coord. | values for units:
10 <L0> 477100.1 7245222.5 672.0 ~ loop coord. | 0 = metres; 1 = feet
11 <L0> 477101.3 7245259.5 678.4 0
12 <L0> 477101.2 7245294.5 684.9 0
13 <L0> 477099.4 7245338.0 686.5 0
14 ~ Hole/Profile Co-ordinates:
15 <P0> 476497.8 7245021.6 672.3 0 0 ~ 0=Collar/Start| 5 numbers: (x,y,z,units,stn)
16 <P1> 476525.7 7245024.5 668.0 0 0 ~ 1=>99=(x,y,z,units) values for units:
17 <P2> 476553.2 7245024.5 658.0 0 0 ~ 2=segment| 5 numbers: (x,y,z,units,segmt length); 1 = feet
18 <P3> 476576.5 7245024.0 655.8 0 75.0 ~|(grid|length| 2az(deg)| dip(deg)| length(m)
19 <P4> 476601.6 7245023.5 666.8 0 100.0 ~|+ve down dip, 3=az(deg) dip(deg) length(ft)
20 <P5> 476625.9 7245025.0 665.6 0 125.0 ~|segmt length| stn: surface PEM (MGS -ve)
21 <P6> 476650.1 7245024.5 666.8 0 150.0
22 <H3> Data Scaled by Coll Area of 650.4300
23 <Tags for comment list: <GEN>, <TA><Stn>, <TX><Stn>, <TY><Stn>, <STZ><Stn>
24 ~ <RDG><Rdg>, <MOD>
25 <Tags for list & plot comments: <SPA><Stn>, <SPX><Stn>, <SPY><Stn>, <SPZ><Stn>
26 <Tags for headings: <HE1>, <HE2>, <HE3>, <HE4>
27 North American Nickel
28 P-008
29
30 P-008
31 August 19, 2015
32 S-COIL Metric Crystal-Master 16.66 150# 10 42
33 #17 9.20 May19,2015z2 150# N 1000 4300 +
34
35 0.000e+00 0.000e+00 1.597e-05 3.193e-05 6.387e-05 1.277e-04 2.555e-04
36 5.189e-04 1.022e-03 2.044e-03 4.087e-03 8.175e-03 1.635e-02
37 $
38 ON XRB 0 A 1940.0 98 256 10 12
39 D4 49.074 -2310.1 41264 41328.7
40 49.07 0.2773 0.2773 0.004442 -0.003059 -0.004113 -0.002263
41 -0.002603 -0.001756 -0.00107 -0.03265
42
43 25N XRB 0 A 1940 98 256 10 32
44 D4 1270.99 -1865.06 35762.5 35833.7
45 1271 -0.1581 -0.1577 -0.05356 -0.008725 -0.003266 -0.001537
46 -0.0002660 0.0005181 0.0006575 -0.2959
47
48 50N XRB 0 A 1940.0 98 192 10 52
49 D4 1433.13 -1711.27 32983.257.7
50 1433 -0.2248 -0.2248 -0.05942 -0.0003389 0.003292 0.002578
51 0.004293 0.008387 0.002182 -1.143
52
53 75N XRB 0 A 1940.0 98 288 10 72
54 D4 1636.21 -1682.48 38737 38826.4
55 1636 -0.2921 -0.2921 -0.05601 0.0003726 0.002262 0.001559
56 0.002201 0.001282 0.0005714 -1.661
57
58 Line 45, Column 65
```

Tab Size: 4 Plain Text



surveys								
id	survey_id	line_id	survey_date	datum	projection	utm_zone	data_format	
							data_units	
7	P-246	NULL	2825-12-03 00:26:51.834163	MGS	1984	UTM	22N	238
6	P-142	NULL	2825-12-03 00:26:33.319472	MGS	1984	UTM	22N	238
5	P-141	NULL	2825-12-03 00:26:03.346828	MGS	1984	UTM	22N	238
4	P-094	NULL	2825-12-03 00:25:28.471354	MGS	1984	UTM	22N	238
3	P-849	NULL	2825-12-03 00:24:56.585166	MGS	1984	UTM	22N	218
2	P-822	NULL	2825-12-03 00:24:49.799994	MGS	1984	UTM	22N	218
1	P-886	NULL	2025-12-03 00:24:06.639254	MGS	1984	UTM	22N	238

em_responses							
id	survey_id	receiver_station_id	component	source_file	station_label	receiver_code	primary_pulse_nt_per_sec
							num_c
1016	4	53	Z	L308EAV.STP	0n	ZH1R	175300
918	4	63	Z	L1308EAV.STP	25m	ZR6	170300
1098	4	53	Z	L408EAV.STP	0n	ZR9	168600
1638	4	53	Z	L308EAV.PDF	0n	ZB18	168233.99693
997	4	53	Z	L1408EAV.PDF	0n	ZR7	16825.40335
981	4	53	Z	L1408EAV.STP	0n	ZR7	165100
940	4	63	Z	L1308EAV.PDF	25m	ZR6	165022.90165
1117	4	53	Z	L408EAV.PDF	0n	ZR9	162931.83268
1099	8	53	Z	L1308EAV.PDF	0n	ZR9	163600

loops									
id	survey_id	loop_point_number	loop_label	easting	northing	elevation	loop_size_x_units	loop_size_y_units	geometry
1	1	1	0	477100.2	724517.0	666.6	600	700	SRID=32622;POINT(477100.2 724517.0)
2	1	1	1	477998.4	7245287.5	667	600	700	SRID=32622;POINT(477998.4 7245287.5)
3	1	1	2	477100.1	7245222.5	672.6	600	700	SRID=32622;POINT(477100.1 7245222.5)
4	1	1	3	477101.3	7245295.5	678.4	600	700	SRID=32622;POINT(477101.3 7245295.5)
5	1	1	4	477101.1	7245294.5	684.9	600	700	SRID=32622;POINT(477101.1 7245294.5)
6	1	1	5	477999.4	7245338	686.5	600	700	SRID=32622;POINT(477999.4 7245338)
7	1	1	6	477100.8	7245366	684.5	600	700	SRID=32622;POINT(477100.8 7245366)
8	1	1	7	477101.8	7245382	678.5	600	700	SRID=32622;POINT(477101.8 7245382)
9	1	1	8	477101.4	7245414	681.1	600	700	SRID=32622;POINT(477101.4 7245414)

receiver_stations										
id	survey_id	station_number	station_label	line_direction	line_index	easting	northing	elevation	distance_along_profile_m	geometry
1	1	0	0n	N	0	476582.9	7245678.5	667.2	0	SRID=32622;POINT(476582.9 7245678.5)
2	1	1	125m	N	1	476531.1	7245678.5	665.8	25	SRID=32622;POINT(476531.1 7245678.5)
3	1	1	25m	N	50	476547.7	7245678	662.1	50	SRID=32622;POINT(476547.7 7245678)
4	1	1	375m	N	75	476578.4	7245678	677.6	75	SRID=32622;POINT(476578.4 7245678)
5	1	1	4 100m	N	100	476597.2	7245675	672.7	100	SRID=32622;POINT(476597.2 7245675)
6	1	1	5 125m	N	125	476621.5	7245675.5	671.5	125	SRID=32622;POINT(476621.5 7245675.5)
7	1	1	6 150m	N	150	476641.4	7245675.5	667	150	SRID=32622;POINT(476641.4 7245675.5)
8	1	1	7 175m	N	175	476675	7245675.5	666.8	175	SRID=32622;POINT(476675 7245675.5)
9	1	1	8 190m	N	190	476697	7245674	660.0	190	SRID=32622;POINT(476697 7245674)

Data Engineering Pipelines

ENVS 5726

- **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

ENVS 5726

- **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

ENVS 5726

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

ENVS 5726

Questions

ENVS 5726



fin