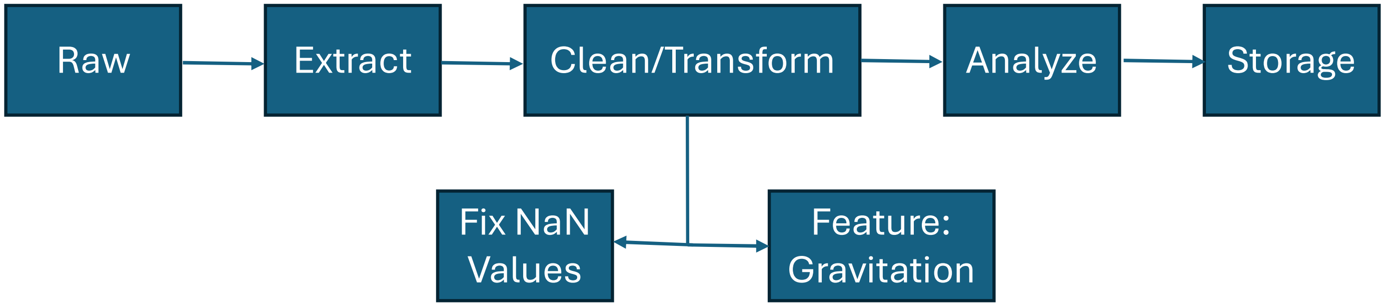
Planet Pipeline

Vincent Huerta

Design

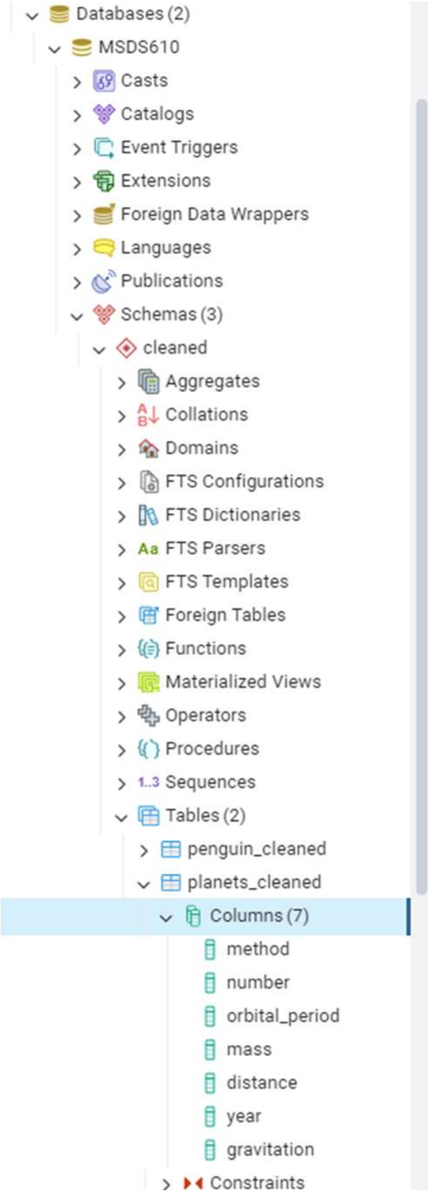


- Batch Pipeline
  - No rational need for instant updates, save resources
- PostgreSQL
  - Text based csv file would be best for organizing and sorting descriptions and traits.

Data Quality – Potential Issues

- Missing Values
  - Missing values would result in entries being removed as nature is unpredictable, even with ML systems.
- Categorical Variables
  - The Number variable represents what type of planet the entry is categorized as, but the categories are man-made based on simple observed patterns, not guaranteed truths. Future entries with no fit can be entered into the Unknown section, but with enough unknowns eventual restructuring might be necessary.
- Hard to identify outliers in chaos
  - Research on planets we cannot reach is limited, outliers can be hard to detect as there is little information and consistency. Even information could be false as data could have gone unchecked, unconfirmed with additional inspection.

PostgeSQL Database


Databases (2)
MSDS610
Casts
Catalogs
Event Triggers
Extensions
Foreign Data Wrappers
Languages
Publications
Schemas (3)
cleaned
Aggregates
Collations
Domains
FTS Configurations
FTS Dictionaries
FTS Parsers
FTS Templates
Foreign Tables
Functions
Materialized Views
Operators
Procedures
Sequences
Tables (2)
penguin_cleaned
planets_cleaned
Columns (7)
method
number
orbital_period
mass
distance
year
gravitation
Constraints