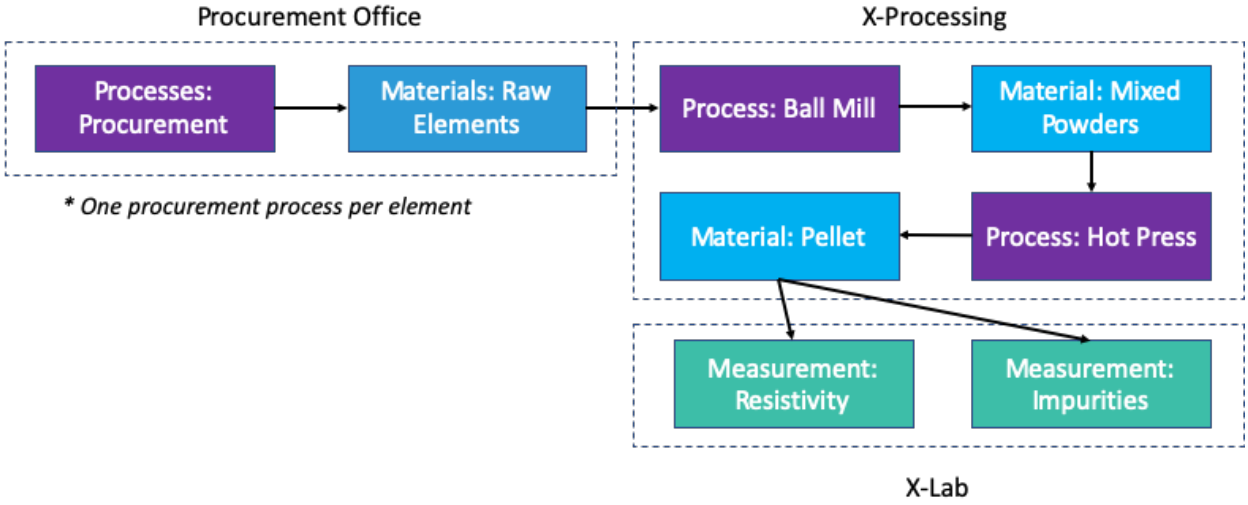
CITRINE INFORMATICS

Data Analysis Engineering Challenge

The Scenario:

Citrine Informatics has recently signed an agreement with X-Materials to use AI in the development of a new semiconductor compound of unprecedented performance. You have been placed as the Data Analysis Engineer in charge of the project.

X-Materials is following a standard procedure to develop their new compound:



One of the major Q1 objectives of this project is to integrate the data of the business units that are part of the material development process:

Procurement Office

- The procurement office is responsible for purchasing the raw materials from vendors.
- They purchase elemental Zn, Cu, and Se from suppliers.

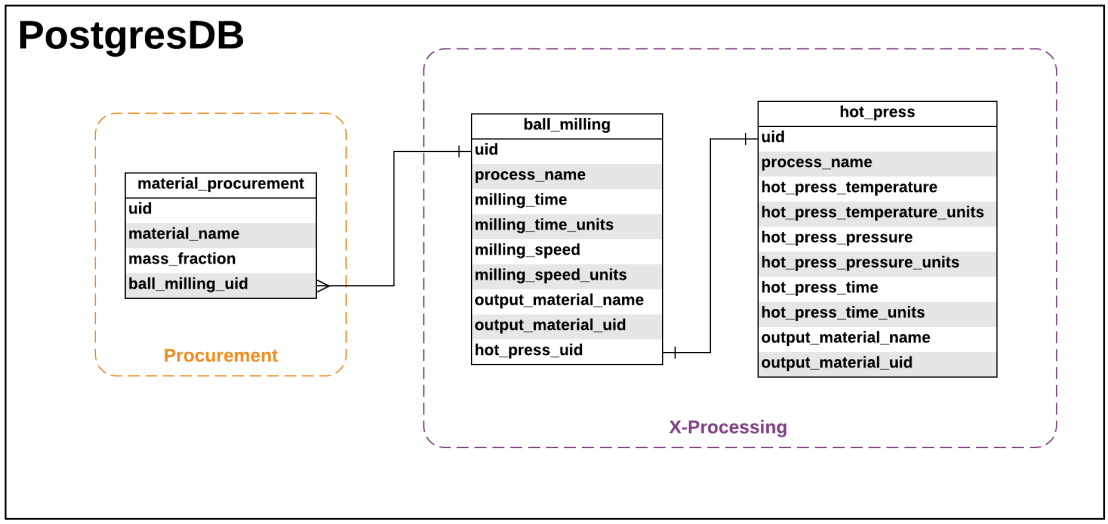
X-Processing

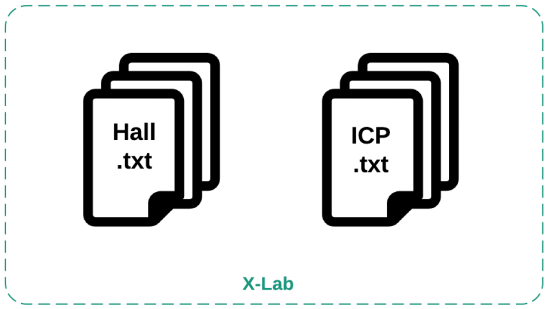
- The raw elements are formed into pellets by the processing lab.
- They mix the raw elements in a high energy ball mill before pressing the resulting powders into circular pellets in a hot iso-static press.

X-Lab

- The X-lab measures composition and electrical resistivity.

Procurement and X-Processing are proficient with their data management. They currently use a Postgres-SQL database to store their relevant data. Within the DB, Procurement has a single table for their entire workflow, and X-Processing has a table for each process they run (ball milling, hot pressure sintering). The X-Lab data, on the other hand, is a data jungle. They have not systemized their data capture and use local txt files to record information for analysis further down the pipeline. Their files contain data of Hall and ICP measurements for materials characterization. X-Materials has provided you with a SQL file representative of the Postgres DB and a set of txt files representative of the laboratory experiments. They have also provided you with the following diagram that illustrates the landscape of their data:





The Challenge: *This challenge should take four hours or less, both parts included. If you find yourself spending substantially more time than that, consider reducing the complexity of your implementation*

I) As the data analysis engineer on this project, you need to integrate X-Materials' siloed data onto a single data management solution that allows their material scientists to generate one master CSV from these disparate data sources whenever new data are added.
You need to execute on the following deliverables in order to have a successful migration:

1. A data pipeline that gathers the historical raw data and pipes it into a master CSV.
2. A way to bring in new data if/when added.
3. Proper documentation
4. A Brief README on how to run your code

Your submission will be evaluated based on:

- Code quality
- Understanding of the raw data
- Ease of use
- Robustness and functionality
- Documentation (both via in-code comments and the README)

II) You are handing off this data pipeline to the IT department of X-Materials. Often, the members of the IT departments who would run such pipelines do not have the domain knowledge of a Materials Science nor they have much first-hand experience working with raw data.

1. Please briefly document, for X-Materials' IT team, why materials data can be particularly challenging data to wrangle.
2. Please briefly document, for X-Materials' IT team, known failure modes of your pipeline, why they may occur, and advice on how they may want to approach these.