Kettle ML Eng. Technical Challenge

First of all, thank you for taking the time to take this challenge. As part of this challenge, we would like to introduce you to a real problem that we are solving at Kettle. By doing so, not only we hope to understand your approaches in solving the problem, but also give you a chance to understand if the work we are doing is interesting and a good fit for you.

Problem Description and Task

At Kettle, we want to accurately predict if fires happen at a given geographical location. Therefore, we deal with various types of geographical data and it is well known that geographical data can be stored in various formats (mainly due to historical reasons), thus are tricky to deal with. For this reason, we want to understand your approach in building a data pipeline and a storage solution for a large geographical data.

The general requirement of the challenge is as follows:

1. We want to retrieve and store a large amount of home parcel data (parcel refers to the plot of land that an owner of a property owns), which is publicly available and may be in multiple formats.
    a. We want to store the data in a uniform format and you get to choose the format.
2. Once stored, we have three use cases for the data.
    a. Shorter analytical queries to retrieve the parcels
    b. Longer analytical queries for running machine learning jobs
    c. Support our real-time interactive mapping solutions.

Given these requirements, we want to see the following:

1. System design of the data pipeline and data storage solution. Be specific about the database solutions that you will be using.
2. **Sample** scripts for the following. These scripts do not have to be long
    a. Schema for data storage. This can be a SQL schema or a JSON schema for other databases. This is up to you.
    b. Script for reading the data downloaded and formatting it to the given schema. Similar to above, the data quality test can be up to you to decide.
    c. Bonus: Test for data quality
3. Some example analytic scripts you can run with your newly created database.
   Can be a SQL query.

Hints and Tips:

1. You can use any parcel data from any source, but an example data and the metadata for NAPA county can be found in http://gis.napa.ca.gov/giscatalog/catalog_xml.asp under “parcels_public” layer. Entity and Attribute Information in the metadata should be very helpful for understanding the data contents.
    a. You can also find and use a smaller data set for smaller counties if that helps
2. Feel free to draw diagrams to support your idea
3. We should be able to run your script from end-to-end
4. Please don’t work on this challenge for more than 5 hours and you may do this in a single sitting or in multiple sittings.