IRisk Lab Data Discovery and Consolidation
Week 9 Report
Oct. 24
Rohit Valmeekam

What Was Accomplished

During Week 9, I shifted the focus to extensive testing of the data pipeline, emphasizing the robustness and scalability of the system. The primary objective was to evaluate the performance of the pipeline under varying conditions, particularly with a substantial influx of input data.

- ❖ Testing the Data Pipeline with Tons of Input Data
- ❖ Data Generation and Diversity:
  - ➢ I initiated the testing phase by generating a large volume of diverse input data, including randomly generated addresses, various property types, and diverse geographical locations. This diverse dataset aimed to simulate real-world scenarios and assess the adaptability of the data pipeline.
- ❖ Input Data Validation:
  - ➢ Rigorous validation of the input data was conducted to identify potential issues or inconsistencies. This step involved checking for data completeness, ensuring the correct format, and verifying the accuracy of the information provided. Any discrepancies found were addressed promptly.
- ❖ Scalability Assessment:
  - ➢ I assessed the scalability of the data pipeline by progressively increasing the volume of input data. This helped identify potential bottlenecks, resource limitations, or performance issues that could arise with larger datasets. The goal was to ensure that the pipeline could handle a substantial workload without compromising efficiency.
- ❖ Logged CPU Usage

Next Steps

- ❖ Find more datasets related to both the AllState Insurance premiums as well as LRO