AI at Scale

Build a unified data delivery platform

fractaloo
INTELLIGENCE FOR IMAGINATION

The Big Picture

A leading insurance provider want to build scalable data preparation tool with the highest level of granularity for group of data scientists. To get there, it needed to ingest, process, and load complex, and deeply nested data. It needed to generate data in a readable format for the end users and develop a unified data delivery platform. The solution also needed to be both extendible and maintainable.

This approach involved several key challenges, such as churning deeply nested XML data in compressed format (~61 TB). The solution also needed to ensure flattening of XML fields with the same number of fields in the processed data irrespective of the schema changes in the source files. The company also sought to automate the entire ETL process as a service.

Transformative Solution

To solve the company's challenges, an approach was taken to collect, process, and consume data.

The collection step involved collecting data from sources. XML data was consumed from HDFS, which was highly nested and complex. Compressed data was read in a scalable fashion (decompression during job runs). From there, the approach wrote custom input formats to handle a variety of input data formats.

The process step involved extracting fields and partitioning. The approach was to iterate through the XML and retrieve field values, generate key-value pairs in a semi-flattened structure, partition the data based on the key (varies at XML level), and generate headers required for output rectangularization.

In the consume step, the data preparation tool was implemented. The approach built a simplified user-interface to prepare the data with the highest granularity. The tool was based on Python Flask, and it provided the ability to choose from a list of available partitions.

The Change

As a result of the engagement, the company gained the ability to:

– Process TBs of data (61 TB) and automate the complete process of ETL.

– Handle compressed and deeply nested data at a large scale.

– Present the most granular details to the data science group in order to streamline their process and improve efficiency.