Data Profiling and Test Automation to Ensure AI-Readiness

> " The Ultranauts team was able to ramp up quickly and get into our organization, seamlessly integrate with our team and understand our business priorities. Ultranauts is unique in terms of their high calibre of quality engineers, and they've been able to get the right processes and structure in place to ensure that we can trust the data being brought into our platforms and systems, as well as the outputs being generated by our models on the other end. "

EVP Data Science

THE DATA SCIENCE TEAM HAS BEEN ABLE TO DELIVER THEIR MODELS

2X FASTER

SITUATION

The client was building an AI-driven brokerage platform from the ground up. The data science and engineering teams were ingesting massive amounts of 3rd party data and rapidly building ML models to power its core platform. Potential new data sources needed to be prequalified to inform data procurement decisions, and once procured, data needed to be cleansed and ingested into the pipeline. The client's Chief Data Scientist recognized that off-the-shelf data profiling and management tools would not be useful given the nuanced judgements required to determine whether a particular data value was "good", and determined that typical outsourced consultants could not handle the task given the domain complexity, the volume of data being ingested, and the speed at which the work needed to be executed.

ENGAGEMENT

Ultranauts deployed a specialized data quality engineering team across two parallel work streams, one prequalifying 3rd party data sources, and the other building structuring and cleaning scripts for the pipeline.

Languages/Frameworks/Tools/Techniques: Data profiling was performed using a toolkit developed by Ultranauts in R. The production system uses Python, a client-specific framework for data acquisition and integration, and data sources in GCS buckets and Bigquery on Google Cloud Platform.

RESULTS

For prequalification, the Ultranauts team worked with business stakeholders to develop usage-specific validation criteria, structure and wrangle the data into analyzable assets, and conduct deep profiling runs to qualify the utility of the data. Ultranauts examined "what-if" scenarios to determine business impact based on data coverage, helping the stakeholders achieve higher ROI on their third party data investment. The team developed a custom toolkit with 120 profiling functions that not only achieved the level of depth and nuance required but also significantly accelerated the findings.

For test automation, the team translated the validation criteria into testable assertions, and automated cleaning pipelines to ensure that all data acquisition routines met requirements.

Ultranauts prequalified a wide range of complex data sets and helped the client make ROI based procurement decisions,

resulting in significant savings of time and money for the client. In one example, the team was able to quantify the number of additional lines of business that could be added in a marketing dataset for a target geography by purchasing new third party data. This required structuring vastly
different data formats to allow a head-to-head comparison, understanding the business intent to define the relevant measures of coverage, and deep profiling of the data sets across multiple dimensions. Ultranauts removed a multi-month burden from the data science and engineering teams, who were able to shift to the higher value tasks of building and deploying core business ML models and meet their business growth goals faster. With Ultranauts' support, the data science team has been able to deliver their models 2X faster.

Client was recently acquired by a market leading insurer. Ultranauts will now be supporting the integration of client's innovative AI platform across a portfolio of business by ensuring consistent data quality across a range of legacy and cutting edge systems.

ultranauts