EmergeFlow
SMALL DIFFERENCE. LARGE IMPACT.

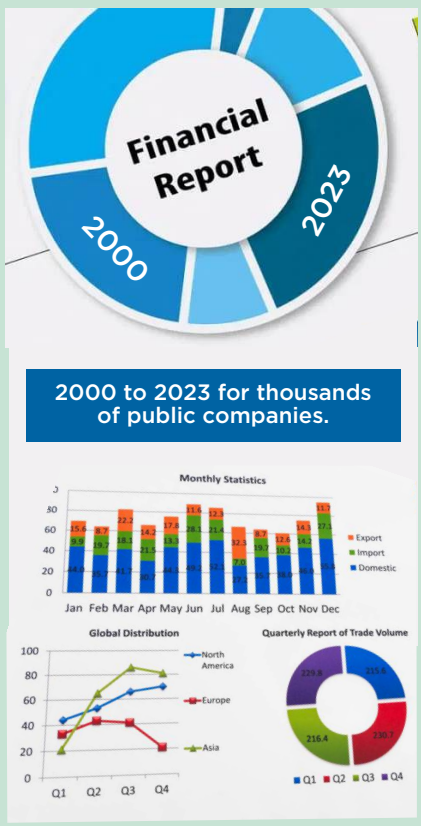
Case study:

Analysing SEC filings using NLP and Automation to understand Corporate tech investments

Client Background

The client is a Dallas-based digital health company specializing in R&D for post-infection recovery software and health insurance.

NEED OF THE CLIENT



The client needed an automated way to analyse over twenty years of SEC filings across thousands of publicly traded companies.

Their main goals were:

- Download and process the historical SEC 10-K reports of all the years from 2000 to 2023 for thousands of public companies.
- Extract the count of Risk Exposure keywords (provided by the clients) from all the downloaded SEC files.
- Connect these risk-data directly with the companies' actual financial data.
- Evaluate risk trends to track how the risks are fluctuating over the years and which sectors were most significantly impacted by changes in the market.
- Run advanced statistical models to test these three hypotheses:
  - H1 (Determinants of Exposure): Firms with greater resources and growth opportunities will have higher Data Center Risk Exposure.
  - H2 (Performance Benefits): Firms with higher Data Center exposure will earn higher subsequent stock returns than peers.
  - H3 (Risk and Pricing): Higher data center exposure increases operational and cyber risks, leading investors to demand a risk premium that manifests as higher expected stock returns.

BEFORE IMPLEMENTATION

The client needed to analyze 23 years of SEC 10-K filings to measure risk exposures across thousands of public companies. Previously, this required manual document downloads and tedious keyword hunting. Merging raw text with financial spreadsheets was nearly impossible, preventing deep statistical analysis of how digital infrastructure impacts market value.

AFTER IMPLEMENTATION

After implementing the automated SEC Risk analysis system, a four-phase pipeline was delivered covering ingestion, NLP-based risk exposure extraction, econometric modeling, and visualization.

Phase 1 established a historical data scraping pipeline that collected all SEC 10-K filings from 2000 through 2023. It includes resumption logic and deduplication to ensure a complete, high-integrity database.

Phase 2 implemented a natural language processing (NLP) engine that reads every ingested filing, systematically identifies 40 distinct technology infrastructure and risk keywords, and writes both raw frequency counts and normalized density scores into a unified dataset. Every record in the dataset carries objective density metrics scaled per 10,000 words, exact mention counts, and temporal markers enabling precise historical analysis across thousands of publicly traded companies. The bag-of-words algorithm was deliberately designed to support rigorous statistical evaluation, translating extensive financial jargon into structured risk and exposure profiles that are directly mappable to external financial identifiers to preserve analytical context.

Phase 3 delivered a sophisticated econometric modeling layer designed for high-stakes testing. This combined Ordinary Least Squares (OLS) regressions and instrumental variable techniques with Random Forest machine learning, all systematically merged with Compustat financial data. This allowed us to rigorously test our three core hypotheses, moving beyond simple correlations to understand the determinants of exposure, performance benefits, and risk pricing.

Phase 4 introduced an automated visualization layer to bring our findings to life. Using scripts for the dynamic generation of trend plots, the system automatically pulls together decades of data and runs the models. This replaces the old way of digging through spreadsheets, delivering actionable graphics that rigorously prove our research outcomes for any given hypothesis.

RESULTS

By running the full suite of scripts, the system successfully transformed mountains of raw text into a unified, high-quality dataset. The modeling confirmed all three core hypotheses.

The pipeline then automatically generated trend plots and charts, making these complex findings easy to digest. We turned thousands of hours of manual labor into an automated process that provides concrete proof of how tech strategies drive corporate value.