

Scaling the Home Office Knowledge Graph by Refactoring the Application

Butterfly Data worked with 6point6 to improve the Home Office knowledge graph, cutting costs and processing time by 85% while boosting stability and scale.

Challenge

The Home Office's Data Assurance Service (DAS) relies on a knowledge graph to provide comprehensive summaries of data quality and lineage statistics across its databases. This graph supports a self-service interface that allows users to query complex relationships and data quality issues.

However, as the volume of raw data grew and new statistics were incorporated, the existing processing pipeline became increasingly inefficient. The system, which updated monthly, began to experience significant delays, rising costs, and frequent instability, hindering timely updates and insights.

To address these challenges and accommodate future data growth, the entire knowledge graph generation process required a comprehensive refactor with a focus on performance optimisation and scalability.

Solution

Butterfly Data, partnering with 6point6, conducted a thorough analysis to identify the primary bottlenecks in the existing system. The investigation revealed that the use of Amazon S3 buckets for storing entity JSON files during processing led to excessive read/write operations, causing substantial overhead as data volumes increased.

To resolve this, the Python-based application was re-engineered to process all JSON files in memory. An efficient in-memory data structure was developed to handle entities with low latency. Additionally, a caching mechanism was implemented for metadata files, ensuring that each file was loaded and processed only once, thereby reducing redundant operations.

The refactor also involved updating numerous class methods to ensure compatibility with the new approach. Enhanced logging features were added to capture subprocess runtimes and provide detailed insights for troubleshooting. Comprehensive testing, including unit and integration tests, was performed to ensure the accuracy and reliability of the refactored system.



66

Butterfly representation in the project team added good value to the project. They worked seamlessly within the project team and were very intuitive with secured technical skills.

Lead Enterprise Data Architect,
6point6

Ready to transform your data?

Book your free discovery call and find out how our bespoke data services and solutions could help you uncover untapped potential and maximise ROI.

hello@butterflydata.co.uk
+44 29 212 00140
butterflydata.co.uk

Get in touch →

Impact

The refactoring efforts resulted in an 85% reduction in processing time and associated costs. The newly optimised pipeline, running on Kubernetes clusters on AWS, demonstrated significantly improved stability, with no aborted runs reported post-refactor.

Furthermore, the process uncovered opportunities for additional improvements and new features, leading to the development of a detailed roadmap. This roadmap equips the DAS team to proactively address future challenges and scale the knowledge graph infrastructure in line with evolving data requirements.



Ready to transform your data?

Book your free discovery call and find out how our bespoke data services and solutions could help you uncover untapped potential and maximise ROI.

hello@butterflydata.co.uk
+44 29 212 00140
butterflydata.co.uk

Get in touch →

Our services



Data quality, governance, and privacy

Improve data quality, compliance and confidence in your data.



Data engineering and cloud adoption

Create scalable cloud platforms with robust data pipelines for multi-source integration.



Analytics and visualisation

Uncover actionable insights and make informed decisions.



Data science and AI

Deploy custom machine learning models and ethical AI to optimise, automate & enhance.