Designing and implementing a data pipeline and data warehouse to store and visualize air quality instrumentation data

Dr Stuart Lacy (Principal Supervisor), Senior Research Software Engineer, IT Services, University of York, stuart.lacy@york.ac.uk
Dr Will Drysdale (Academic Supervisor), Postdoctoral Research Associate, Wolfson Atmospheric Chemistry Laboratories, University of York, will.drysdale@york.ac.uk

The Wolfson Atmospheric Chemistry Laboratories (WACL) represent a premier UK research institution, integrating the expertise of approximately 70 researchers across ten academic groups. The laboratory's mission—investigating gas and aerosol phase atmospheric processes—is underpinned by a vast infrastructure of 80-100 instruments. These assets are deployed globally, from long-term monitoring stations to intensive field campaigns, generating an expansive and heterogeneous dataset totaling ~16TB.

While preliminary efforts have consolidated these raw outputs into on-premises storage, the diversity of data formats continues to present a significant barrier to efficient analysis. This internship seeks to bridge that gap by advancing the development of a sophisticated Research Data Management (RDM) strategy. The primary objective is to evolve an existing prototype into a robust, scalable data pipeline that transforms raw instrumental output into a structured, queryable architecture suitable for both high-level research and real-time diagnostic monitoring.

The project will leverage a high-performance, hybrid-cloud infrastructure to ensure data integrity and accessibility:

- Pipeline Orchestration**:** Utilizing Nextflow and Globus to manage scalable data transfers and automated workflows across High-Performance Computing (HPC) environments.
- Data Processing & Reproducibility**:** Implementing R-based processing scripts encapsulated within Apptainer containers to guarantee computational reproducibility across different environments.
- Storage & Querying: Maintaining a dual-layered storage approach using on-premises servers and AWS (S3). The cloud layer will leverage AWS Athena to provide a serverless SQL interface, enabling researchers to perform complex queries across the entire data warehouse.
- Visualization**:** Developing live Grafana dashboards to provide instrument maintainers with real-time telemetry, facilitating proactive maintenance and reducing data gaps.

By optimizing the laboratory's Research Data Management framework, this project will significantly accelerate the research cycle. Researchers will benefit from near-instantaneous data retrieval, while technical staff will gain enhanced oversight of instrument health. This integrated pipeline not only preserves the long-term value of WACL's data assets but also ensures the laboratory remains at the forefront of atmospheric data science.