IT Infrastructure for a Data Science Campus

# Introduction

The modern economy provides both the challenge of measuring fast-evolving forms of economic activity, and the opportunity to exploit huge amounts of new data and information to help policymakers, researchers and businesses. The Data Science Campus (the Campus) at the Office for National Statistics (ONS) has been created to respond to this challenge and will act as a hub for the whole of the UK public and private sectors to gain practical advantage from the increased investment in data science research and capability building. Our goal at the Data Science Campus is to explore how new data sources and data science techniques can improve our understanding of the UK’s economy, communities & people. Developing a data science platform and infrastructure that meets these goals whilst ensuring data governance and security poses significant challenges. We outline the services provided within the ONS and the Campus to address these needs.

# ONS Data Service

The ONS corporate network is undergoing significant transformation. It is a highly secure area containing standard corporate IT services such as Email, Skype for Business, and SharePoint. IT services are separated and secured through secure zoned networks. The ONS Data Service consists of various network zones that integrate together providing the business with services to securely ingest data into a distributed data lake. This contains the bulk of the business’s sensitive data. The service also provides tools for the preparation and exploration of data, as well as production and export pipelines. Users access the platforms and tool sets from the corporate network using Virtual Desktop Infrastructure (VDI).

The secure and robust implementation of the corporate infrastructure allows data science techniques to be applied to sensitive data; however due to security constraints, it also restricts the libraries, models and tools that can be used. This can limit innovation and constrains the ability for data scientists to implement and experiment with new tools and develop new techniques. The Data Science Campus network (DSCN) has been created as separate infrastructure to provide users with IT services and tool sets required to investigate more advanced techniques and produce the next generation of statistics.

 Corporate network vs. demilitarized zone (DMZ)

**CORPORATE NETWORK**

Corporate IT Services

Trusted Data Zone (TDZ)

Data Access & Exploration Platform

Zoned, isolated networks

**SENSITVE DATA**

**CAMPUS NETWORK**

Innovation Platform

Isolated from corporate network (DMZ)

VLAN isolation

Available to off-network devices

**NON-SENSITVE DATA**

# ingesting data

Data is central to the work at ONS. Data enters the business from a multitude of sources in both structured and non-structured forms. The security, access control and regulation of data is of the upmost importance. The Trusted Data Zone (TDZ) was created to securely ingest this data for use in the ONS Data Service platform. The TDZ contains tools to handle data ingestion, including Apache NiFi and API Gateways. The zone also contains a multi Anti-Virus scanner to ensure that the data is free from malware and viruses before entering the network. Data on the Campus network is non-sensitive, and the network is designed so data cannot enter the corporate network. The data is virus checked before being stored in a manner that is appropriate for its end use.



# the data science platform

The DSCN sits in the Demilitarized Zone (DMZ) in complete isolation from the corporate network on dedicated hardware. The DSCN is accessible from the corporate network and external networks. Users can access and use a variety of tools and can develop their own data systems as required, including many that would be restricted on the corporate network. The system is designed to encourage innovation and experimentation. It has been used to develop techniques and tools that can later be migrated to the corporate data service using continuous integration pipelines.

# example Projects

The data science campus network has been utilised to deliver various projects, these will be outlined and include:

* Access to Services - revealing travel information regarding access to health, social and other services. Uses the Campus propeR tool that uses open-source transport data to analyse multi-modal travel in the UK.
* Identifying household gardens - accurate estimate of the green space within UK household gardens.
* Identifying emerging trends from patent data - patents and other technical literature have key terminology trends identified, which may inform business and government decisions regarding new technologies. The analysis includes when and where terminology usage occurs, considered both nationally and internationally. This builds on the Campus PyGrams pipeline that allows information retrieval from large document collections.
* Turning free-txt lists into hierarchical datasets - using optimus, a processing pipeline which can automatically group and generate hierarchical labels for each group to structure the data
* Mapping the urban forest - generating vegetation density maps by for the road network of an entire city. The developed system is built on recent advancements in the field of deep learning for semantic image segmentation.
* Re-using the approach - the Campus is partnering with the National Institute of Statistics Rwanda (NISR), supporting them to deliver a similar IT infrastructure for a new Data Hub in Rwanda.
* Training and development sandpit – to develop data science capability within ONS and Government wide, with resources to teach R, Python, Spark, Git, Natural Language Processing and machine learning.