Developer and method services in the UN Global Platform

Keywords: algorithms, statistics, serverless, cloud, infrastructure, data science, platforms.

1. Introduction

Governments and their statistical offices face global challenges. A good example of a challenge at this scale is the development of indicators to support the Sustainable Development Goals in an age in which data is fast, big and heterogeneous. Similarly, most official statistics are published to agreed international standards. This situation has led to statistical offices making overlapping and, at a global scale, redundant investments, while not leveraging their combined resource pool of knowledge, infrastructure and capital.

1.1. CSPA - history

In the domain of technology this problem has been previously approached by initiatives such as the Common Statistical Production Architecture [1] (CSPA). Namely: can statistical offices develop and build reusable IT components that solve common business problems? The advantages of such an approach are: a) it drives a reduction in cost by building once and using many times, b) it promotes standardisation by investing in fewer but better understood components, c) it delivers faster adoption of solutions by having a standardised form of deployment. However, at its inception CSPA has met many obstacles for adoption. Some of these include: a) the technological solutions were more complex in nature than those in use by the local statistical offices themselves, b) the statistical community did not have the capability to fully utilise technologies that facilitate reuse (such as containers, cloud or restful APIs), c) although the reuse approach was agreed in principle, it was not matched by joint investments.

CSPA was informed by a set of best practices inherited from Service Oriented Architecture (SOA) and microservices. However, examples of developing a consistent practice of production of CSPA style services are scarce. Our development of a global methods service fundamentally facilitates the production, deployment and consumption of CSPA-compatible services, but reducing the high effort and high cost technical support required previously.

1.2. Technology advances post 2010

In the last couple of years a number of technologies that can support reuse and portability have matured: a) containers [2] (such as Docker and Kubernetes) have become commonplace for enterprise application storage, deployment and scaling, b) public and hybrid clouds [3] are the de facto corporate IT infrastructure and governments are adopting them, c) more solutions are

available to automate the deployment, maintenance and testing of containers, applications and APIs. The maturation of these technologies provides large opportunities to the statistical community, while the community also risks not developing skills while they become more widely adopted as standard IT practice.

Containers have been previously identified as one of the solutions to the portability of statistical services. They allow an IT solution to be built and stored including its complete execution environment and dependencies. One single machine or virtual machine can run many containers. Their self-containment makes them ideal for sharing by copying from one environment to another, as well as for scaling them by distributing them into many machines.

Public cloud allows organisations to acquire infrastructure on demand hosted by cloud providers in off-premises data centres. Although this practice is not new, the maturity, accessibility and variety of the services currently provided is at a new peak. The re-focusing of IT infrastructure towards cloud creates a change in culture: infrastructure can be disposed off when not needed. This has led to the emergence of new patterns for services, such as Function-as-a-Service [4] (FaaS) and Platform-as-a-Service (PaaS). Statistical offices can now “spin-up” infrastructure for a single purpose (for example, a public-private partnership, data collaborative or hackathon), and dispose of it (and its associated costs) at the end of the project.

Applications that make use of containers, cloud, and ‘serverless’ or FaaS principles are becoming more common. These take the focus away from developers, architects and data scientists tinkering with infrastructure, and focus their efforts into the source of value: developing computing logic with business value. For example, AWS Lambda and Google Functions abstract away the main infrastructure concerns and allow developers to focus on coding functions… which by default will be available, accessible, scalable and fault tolerant.

2. UN Global Platform: developer and method services

The Statistical Commission of the United Nations created the Global Working Group on Big Data [5] in its 45th session. The aims of the working group are to explore and exploit the opportunities of big data for the statistical community and build international capability. The UN Global Platform [6] is being developed under the Working Group in order to provide access to UN countries to the latest technologies and explore the possibilities provided by emerging trends.

As part of the prototyping and development of the Platform, we provide access to users in the statistical community to several services. Each service is cloud-native, has been developed in partnership with a technology company and has its own site and access controls. Two of the most mature services we have developed are:

1. The developer service. This service provides users an internationally accessible collaboration environment for data science and development of code.
2. The methods service. This service allows users to develop, publish and consume algorithms and statistical methods.

These services interplay with each other. The developer service can be used to build new algorithms, methods and machine learning models. These can then be published in the methods service as easy-to-consume microservices; which can, in return, be consumed from the developer service.

2.1. Methods service

The UN Global Platform methods service has been developed in partnership with Algorithmia since April 2018. It takes the core of the commercial Algorithmia site and redeploys these elements into our own, globally-accessible, cloud infrastructure. The service allows statisticians and data scientists to write and publish callable algorithms in a number of programming languages (for example, mathematical functions, statistical methods or machine learning models) without having to deal with IT aspects such as dependency management, container building or scaling. Each algorithm is compiled by the service and turned into a container, which is accessible by a published API. In a sense, the service automates the production of 'shared' CSPA services. Additionally, the service can host machine learning models such as neural networks or decision trees, and publish functions provided by the model as APIs.

As of October 2018 the service has over 30 users of 10 organisations from different countries. To date we have demonstrated:

- We can publish statistical methods and mathematical functions and consume them from anywhere.
- We can scale the computing power of the platform to meet demand by implementing asynchronous calls that make use of scaling processing across nodes / containers.
- We can develop and publish complex machine learning models such as the Urban Forests pipeline developed by the UK Data Science Campus.
- We can build on existing algorithms by including them in new algorithms.

Our main learnings have been:

- The service can significantly speed up development and testing of algorithms and models by taking away the 'devops' overhead (an existing machine learning model can be deployed and shared by one person in hours rather than days).
- The service can make it easier to implement algorithms in production by providing versioning and speeding up prototyping.
- Microservice-stylethinking can create granular statistical services that can be composed and extended by other services.

2.2. Developer service

The UN Global Platform developer service has been developed in partnership with Anaconda since July 2018. It is a deployment of Anaconda Enterprise within our cloud infrastructure, and available globally. The service allows users to collaborate in shared 'projects', where they can upload data, manage packages in Python and R, and work in Jupyter Labs notebooks. Access to data within projects is limited to members of the project.

The main purpose of providing the service is so that users can explore new (shared) datasets securely, develop analytical pipelines to exploit those datasets, as well as to produce new algorithms that can be published in the methods service. The cloud deployment allows the computing power and storage of the service to be easily scaled depending on user demand (for example, more concurrent users will require more cores to be provisioned).

As of October 2018 the service has over 20 users from 9 organisations from around the world. To date we have demonstrated:

- We can securely share data within projects for international partners.
- We can develop and test methods and algorithms, and manage their dependencies.
- We can consume methods, algorithms and machine learning models from the RESTful APIs of the methods service from the developer service (with both synchronous and asynchronous calls).

Our main learnings are:

- Shared developer practices internationally are hard - more guidance is needed.
- Security and cloud, although cliche, need communicating and a culture change.

References

[1] UNECE, 2015. Common Statistical Production Architecture. [Online] Available at https://statswiki.unece.org/display/CSPA/. [Accessed 15 October 2018].

[2] Google, 2018. Containers at Google. [Online] Available at https://cloud.google.com/containers/. [Accessed 15 October 2018].

[3] Jonas de Muro, 2018. What is Public Cloud? [Online] Available at https://www.techradar.com/uk/news/what-is-public-cloud. [Accessed 15 October 2018]

[4] Bowei Han, 2017. An Introduction to Serverless and FaaS. [Online]. Available at https://medium.com/@BoweiHan/an-introduction-to-serverless-and-faas-functions-as-a-service-fb5cec0417b2. [Accessed 15 October 2018].

[5] UN Statistics Division, 2018. Big Data UN Global Working Group. [Online] https://unstats.un.org/bigdata/. [Accessed 15 October 2018]

[6] UN Global Platform, 2018. UN Global Platform. [Online] https://www.officialstatistics.org/. [Accessed 15 October 2018].