Strategies for migrating workflow manager in live, clinical genetic diagnostic service – minimizing risk, maximizing quality

Genomics England | Genie

Błażej Szczerba
Presenter, Ardigen*

Ardigen

Abstract

Migrating a live, clinical genetic diagnostics service means that minimising risk and divergence during the migration are top priority.

Genomics England provides whole genome sequencing diagnostics to the Genomic Medicine Service (U.K), a free at the point-of-care, nationwide, genomic diagnostic testing service, with ambitious targets of processing 300,000 samples by 2025. Currently, all clinical bioinformatics is processed using a clinical-standard certified, internally developed workflow engine (Bertha). We are migrating to a new solution (known as Genie, based on Nextflow and Nextflow Tower) which combines off-the-shelf products with custom functionality, so we can focus on our core mission to enable equitably accessed, genomics medicine for all.

Genomics England are currently utilizing an in-house workflow manager for genome sample processing, but we are now looking to migrate to a different workflow manager. To determine the optimal approach for this migration, our foremost concern is risk mitigation, minimizing time to migrate and that the quality of our processes remains at its highest (working in small iteration and minimizing divergence). After careful evaluation of three distinct migration strategies, we have decided to adopt a 'lift and shift' strategy, which involves transferring our existing setup to the new system as is, to minimize disruption and maintain continuity. Furthermore, the migration process will be executed following an agile and iterative methodology, accompanied by a robust automated testing framework. This approach will help us continuously assess and mitigate any potential risks, ensuring a smooth and reliable transition while preserving the quality of our workflow management.

1. Introduction

Innovating Workflow Management: Our Transformation Journey

Bertha overview

Bertha, our current in-house workflow manager, is custom-designed to meet the specific requirements of the 100k Genomes project. It has been implemented entirely in Python, allowing seamless integration with a range of Python-based tools and providing the flexibility essential for our workflow.

Notably, Bertha is tightly coupled with our on-premise infrastructure, ensuring efficient and secure data management and computation resources that are instrumental to our project's success.

While its monolithic business logic has effectively served the needs of our project, we recognize that the evolving landscape of workflow management tools offers solutions with even more advanced capabilities.

Migration reasoning

As we continue to evolve and expand our clinical genome processing offering for the National Health Service (NHS) in England, we're exploring alternative solutions on the market.

This decision is driven by the recognition that while Bertha has effectively met our current needs, the field of workflow management tools has seen significant advancements in recent years, offering enhanced features and capabilities.

By considering these alternative tools, we aim to find a solution that aligns even better with our evolving requirements, ensuring increased efficiency in context of execution environment and easy adaptability for future use cases.

The 100,000 Genomes Project

Why Nextflow ?

Nextflow supports multiple execution environments, making it adaptable to a range of computing setups, from local machines to cloud platforms. Nextflow also offers Nextflow Tower, a standalone workflow execution manager with a web interface. Nextflow has gained significant traction within the bioinformatics community, resulting in improved support for our requirements, libraries, best practices, and guidelines.

What makes Nextflow especially appealing is its flexibility and module creation capabilities. It allows users to create and integrate customized components, making it easy to build complex workflows with intricate logic without extensive coding. This makes it an excellent choice for projects with demanding computational needs and complex workflow requirements.

Genie should help us support newer use cases quicker, across different infrastructures such as cloud, offers GA4GH WES APIs and uses a standard workflow definition language.

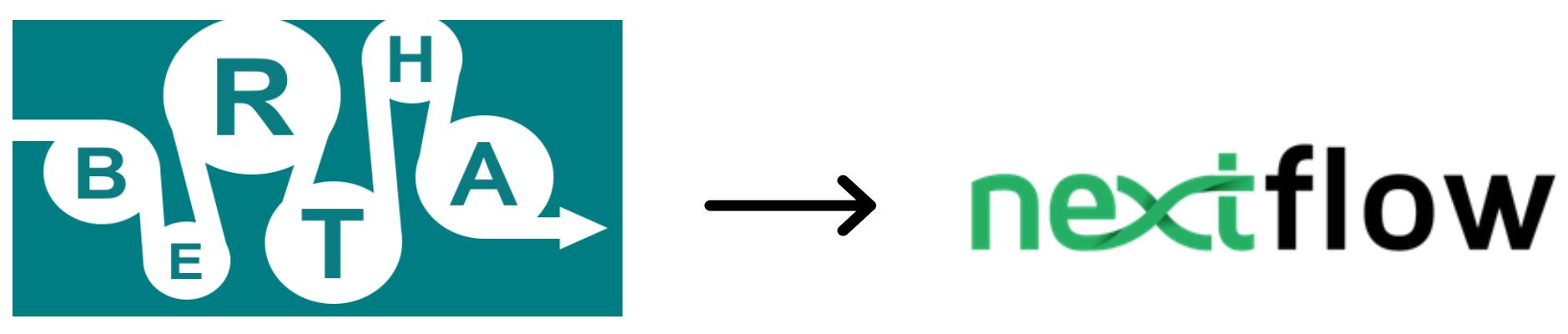
2. Strategies

We explored three approaches for migrating from Bertha to Nextflow.

Strategy A – Rewrite from ground up

Rewriting workflows from the ground-up would allow us to develop an entirely new solution with *state-of-the-art* features of modern workflows managers as Nextflow.

Pros: This solution allows for separation of business logic code from workflow management. Starting from the scratch (or in shallow refactoring) Nextflow workflows and modules allows to great resource management, process parallelisation, easily access to intermediate results



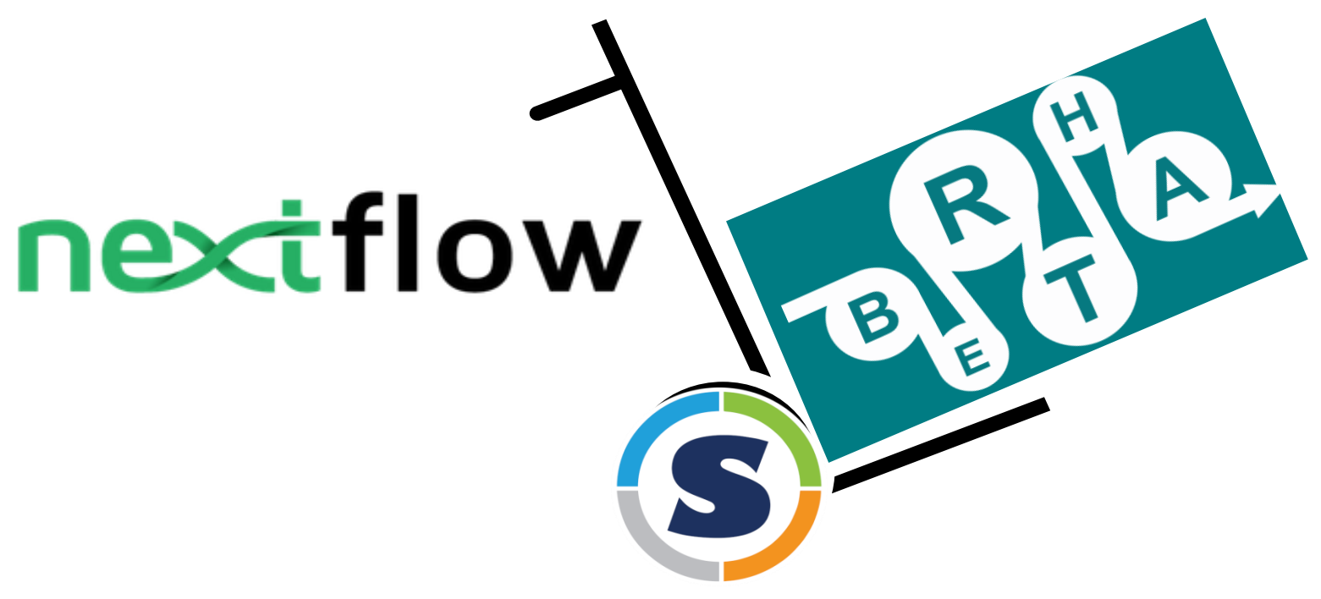
Cons:
Highest risk of potential divergence, need to code freeze in Bertha

Strategy B – Lift and shift

Encapsulate Bertha's components in Nextflow processes to not change actual code of components

Pros: This strategy reduces the risk of divergence in the transition period. While the code running in Bertha and Genie are the same during the transition period, the strategy doesn't prevent doing minor changes in the components that may simplify the migration.

This ensures that new requirements, optimisations and refactors are immediately available in the both environments.

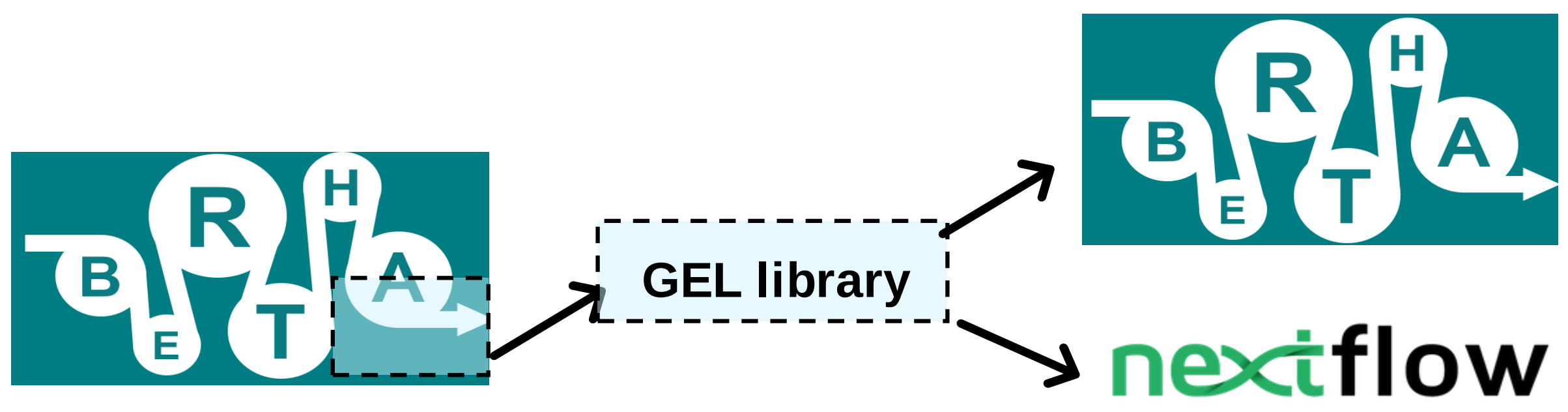


Cons:
We stay with monolithic business logic longer than we want so technical debt will be growing

Strategy C – Extract business logic

Create GEL Libraries – python packages, an independent, workflow-agnostic repositories for housing business logic extracted from Bertha components. It ensures that critical calculations and outputs remain consistent across various workflow managers.

Pros: It offers minimizing the risk of behavioural inconsistencies during parallel development of Bertha and Nextflow pipelines, promotes easy integration with other workflow managers, easy sharing logic between multiple components and simplifies system maintenance.



Cons:
There is extra time needed to extract logic to independent library. To create and maintain coherent interfaces to use in Bertha and Nextflow can be very demanding. Challenges include the need for refactoring during component extraction and interface design considerations

3. Migration priorities

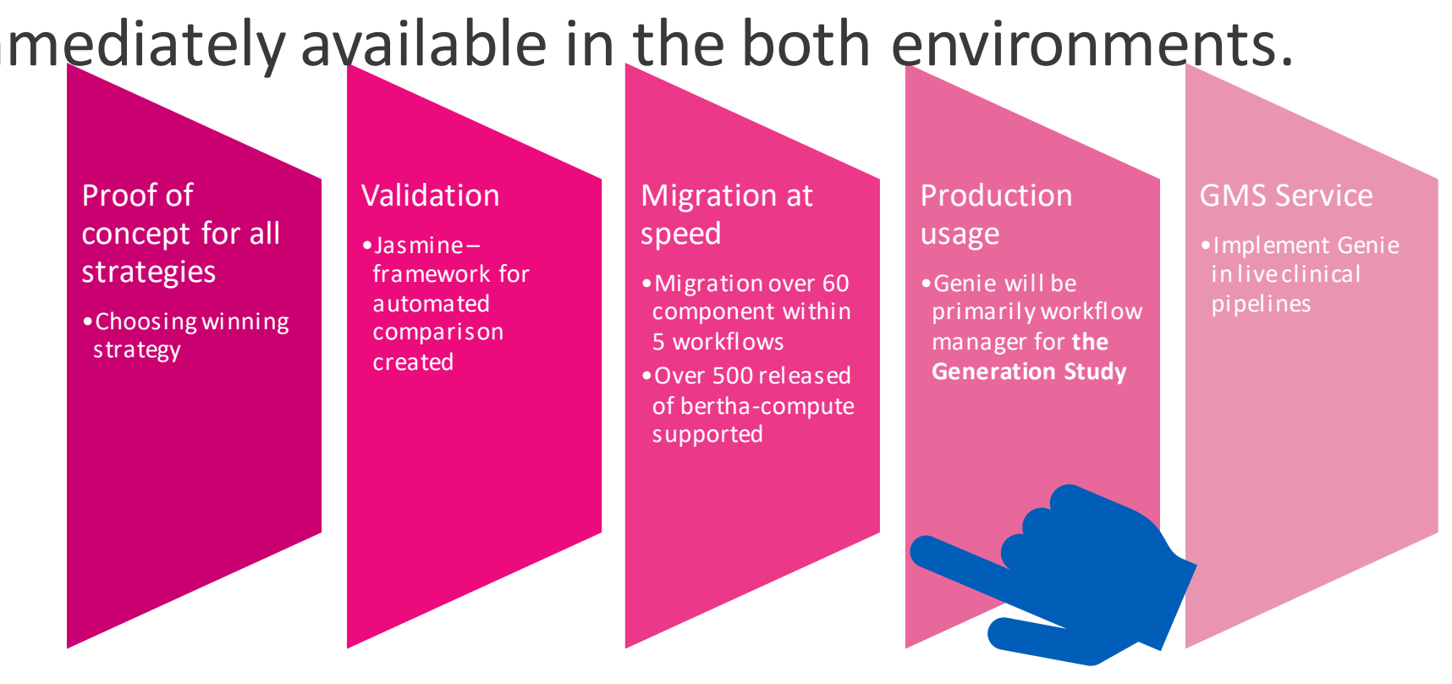
Migrating a live, clinical genetic diagnostics service means that minimising risk and divergence during the migration are top priority.

❖ Minimise risk of divergence between Bertha and Nextflow workflows during migration process.
❖ Minimise patient, data integrity, configuration, clinical content errors, incorrect data validation and data errors, and testing risks.
❖ Work in small iterations, delivering value regularly and fast.
❖ Deliver migration on time for The Generation Study

4. Winning strategy

Lift and shift Strategy B
As the initial phase of our migration process, we opted for a 'Lift and Shift' approach. This strategy preserves the existing code of the components in the main branch without making any modifications; instead, it utilizes the components from the current Bertha Compute image.

This should reduce the risk of divergence in the transition period. **While the code running in Bertha and NextFlow is the same during the transition period**, the strategy doesn't prevent doing minor changes in the components that may simplify the migration. This ensures that new requirements, optimisations and refactors are immediately available in the both environments.


Proof of concept for all strategies
- Choosing winning strategy

Validation
- Jasmine – framework for automated comparison created

Migration at speed
- Migration over 60 component within 5 workflows
- Over 500 released of bertha-compute supported

Production usage
- Genie will be primarily workflow manager for **the Generation Study**

GMS Service
- Implement Genie in live clinical pipelines

5. Next steps

Automated validation framework
We have implemented a novel, automated comparison test framework - **Jasmine** so that we have a validation benchmark for our new pipeline. A comprehensive examination of this topic can be found within a poster created by Ricardo Ramirez, titled "**Migrating a complex workflow to Nextflow with a containerised replica of the production environment**".

Refactoring pipelines
Following the successful completion of the Lift and Shift migration, our next phase involves optimizing our pipelines to fully leverage all of Nextflow's capabilities and achieve greater independence from bertha-compute.
For further details, please refer to Luke Paul Buttigieg's poster, titled **"Making Genomics England's clinical genomic workflows future-proof: refactoring strategy"**

Ardigen, **an AI CRO, assists** companies in tackling large-scale bioinformatics projects by offering expert guidance and technology solutions. We help organizations create and transition complex bioinformatics pipelines, ensuring smooth data integration and analysis. This allows companies to make significant progress in their research and development efforts.

More information: genomicsengland.co.uk

blazej.szczerba@ardigen.com