BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
https://excel.fit.vutbr.cz

# A System for Distributed Computation Task Management in Containers

Bc. Kristián Kováč*

**Abstract**

Securely analyzing sensitive genomic data across multiple institutions is technically complex, due to strict security rules, complicated access management, and long-running workflows. This project solves these problems by creating a federated system that runs workflows using Snakemake, together with TES APIs and Funnel. Users are authenticated and authorized with OAuth and GA4GH Visa tokens; these tokens are automatically refreshed even during very long computations. The system runs in Kubernetes, uses Celery to run tasks reliably, and provides a dashboard to manage and track workflows easily. The resulting federated architecture facilitates secure, effective cross-institutional collaboration and advances practical standards for genomic data processing.

*xkovac61@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

## 1. Introduction

International collaboration has become increasingly important in genomic research. Nevertheless, efficient and secure analysis of sensitive genomic datasets across institutional and national boundaries remains a considerable technical challenge. Genomic data frequently include highly sensitive patient information, necessitating stringent protection against unauthorized access during collaborative analyses. Additionally, computational workflows in genomics often require hours or even days of uninterrupted execution, complicating the secure management of authorization credentials, as authentication tokens typically expire during extended computations. Addressing these issues requires secure, scalable, and intuitive technical frameworks that can reliably run prolonged workflows without interruption or security vulnerabilities.

Initiatives such as Beyond 1 Million Genomes (B1MG) and 1+ Million Genomes (1+MG) have been established to standardize data-sharing policies, governance, and collaboration strategies internationally. Although considerable progress has been made toward policy standardization, practical technical implementations addressing challenges—such as federated computation, stable authorization management, and continuous execution during lengthy analyses—remain limited.

In this project, a federated workflow execution platform designed specifically to overcome these technical gaps is presented. A secure infrastructure enabling efficient cross-institutional computation of sensitive genomic datasets has been developed. Additionally, an automated mechanism for token refreshing has been integrated to maintain continuous authorization throughout prolonged computations. To further enhance usability, an intuitive dashboard for simplified workflow initiation, real-time monitoring, and debugging is provided. Kubernetes orchestration combined with Celery task management has been employed to ensure scalable, reliable, and secure operation, significantly advancing technical capabilities critical to collaborative genomic research aligned with the objectives of B1MG and 1+MG initiatives.

## 2. Genomic Data Infrastructure services

Genomic Data Infrastructure (GDI) provides integrated components for secure discovery, authorization, and analysis of sensitive genomic datasets in collaborative environments [1]. Users interact with the infrastructure primarily through the User Portal, which includes a Data Catalogue to discover relevant datasets and an Access Management module for requesting data access. Resource Entitlement Management System (REMS) manages these requests

and issues GA4GH Visa JWT tokens that encode detailed data-use permissions [2]. Authentication and authorization are securely validated through the Life Science Authentication and Authorization Infrastructure (LS AAI) in collaboration with REMS, ensuring controlled access to sensitive genetic data [3].

Genomic datasets are securely stored in Sensitive Data Archives (SDA) that restrict external data access [4]. Researchers utilize Beacon, a federated querying service, for secure and privacy-preserving exploration of data availability, without directly accessing sensitive data. For conducting detailed analyses, compute backends are authorized to retrieve data securely from SDA via standardized HTSGET protocol, rigorously maintaining granular authorization and data protection [5].

## 3. Federated architecture

The Genomic Data Infrastructure (GDI) employs a federated architecture consisting of multiple independent nodes, each securely handling sensitive genomic data computations. Within each node, the Workflow Executor coordinates user-defined computational workflows, passing tasks to a Funnel-based Task Execution Service (TES) for secure, containerized execution [6]. Sensitive genomic datasets remain secured within each node's dedicated SDA, accessible only through strictly authorized internal service requests. This decentralized architecture ensures data security and facilitates effective collaboration, enabling workflows across institutional or national boundaries without exposing sensitive data directly.

## 4. Workflow executor

The workflow executor is designed for managing computational tasks reliably and efficiently, integrating Python's Celery framework and Snakemake workflow orchestration [7]. The executor provides a REST API, allowing users and external applications to submit workflows and monitor progress. Incoming workflow requests are transformed into executable tasks by the Task Builder and placed into Celery's distributed queue system. Celery workers then manage these tasks concurrently, with Snakemake orchestrating individual workflow steps, tracking jobs, resolving dependencies, and automatically refreshing authorization tokens for long-running computations. This combination ensures robust, scalable, and uninterrupted workflow execution, supports graceful updates with minimal disruption, and simplifies the operation, monitoring, and maintenance of complex genomic analysis workflows.

## 5. Snakemake

To extend Snakemake for secure genomic data workflows, custom modifications and features were implemented. First, a dedicated Python library was developed to handle OAuth token management, enabling Snakemake to securely pass authorization tokens to computational executor backends. Accordingly, a custom TES plugin was created using this library, ensuring secure authentication during task submission and execution.

Additionally, Snakemake was enhanced to recognize SDA storage paths directly within workflow definitions (Snakefiles). Users define input data sources using these SDA paths, which Snakemake does not attempt to resolve directly (unlike typical storage providers). Instead, the executor plugin transparently forwards these paths into the TES requests using the specialized `sda://<path>` notation supported by the Funnel backend. This approach ensures each SDA path corresponds to the local SDA instance of the specific federated node where the computational task is executed, maintaining strict security boundaries and precisely controlled access to sensitive genomic data during federated workflow execution.

## 6. Conclusion

In this project, a secure, federated workflow platform was successfully developed, tailored specifically for sensitive genomic data analysis in cross-institutional collaborative environments. Leveraging Snakemake, custom TES executors, OAuth-based token management, and standardized authorization via GA4GH Visas, this solution addresses security, authorization, and long-running workflow execution issues effectively. The implemented federated infrastructure, coupled with intuitive user interfaces and scalable Kubernetes and Celery integration, significantly enhances real-world capabilities for international genomic collaborations. Through this work, substantial progress was achieved toward securely realizing goals outlined by initiatives such as B1MG and 1+MG, ultimately facilitating more practical, secure, and effective genomic research partnerships across institutions.

## Acknowledgements

I would like to thank my supervisor RNDr. Marek Rychlý, Ph.D. and consultant RNDr. Lukáš Hejtmánek, Ph.D. for their help.

## References


[1] E Pascucci, F A Causio, G Pesole, M Chiara, M Morelli, G Tonon, F Nicassio, R Pastorino, G E Calabrò, and S Boccia. Progressing towards personalised medicine: the genomic data infrastructure (gdi) project. *European Journal of Public Health*, 34(Supplement_3):ckae144.1956, 10 2024.

[2] Craig Voisin, Mikael Linden, Stephanie O.M. Dyke, Sarion R. Bowers, Pinar Alper, Maxmillian P. Barkley, David Bernick, Jianpeng Chao, Mélanie Courtot, Francis Jeanson, Melissa A. Konopko, Martin Kuba, Jonathan Lawson, Jaakko Leinonen, Stephanie Li, Vivian Ota Wang, Anthony A. Philippakis, Kathy Reinold, Gregory A. Rushton, J. Dylan Spalding, Juha Törnroos, Ilya Tulchinsky, Jaime M Guidry Auvil, and Tommi H. Nyrönen. Ga4gh passport standard for digital identity and access permissions. *Cell Genomics*, 1, 2021.

[3] M Linden, M Procházka, I Lappalainen, D Bucik, P Vyskocil, M Kuba, S Silén, P Belmann, A Sczyrba, S Newhouse, L Matyska, and T Nyrönen. Common elixir service for researcher authentication and authorisation [version 1; peer review: 3 approved, 1 approved with reservations]. *F1000Research*, 7(1199), 2018.

[4] Dietmar Fernández-Orth, Audald Lloret-Villas, and Jordi Rambla De Argila. European genome-phenome archive (ega) – granular solutions for the next 10 years. *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 4–6, 2019.

[5] Dylan Spalding, Juan Arenas Marquez, Salvador Capella-Gutierez, Tommi Nyrönen, Bengt Persson, and Serena Scollen. Data standards and the european genomic data infrastructure, 6 2023.

[6] Alexander Kanitz, Matthew H. McLoughlin, Liam Beckman, Venkat S. Malladi, and Kyle Ellrott. The ga4gh task execution application programming interface: Enabling easy multicloud task execution. *Computing in Science & Engineering*, 26(3):30–39, 2024.

[7] F Mölder, KP Jablonski, B Letcher, MB Hall, CH Tomkins-Tinch, V Sochat, J Forster, S Lee, SO Twardziok, A Kanitz, A Wilm, M Holtgrewe, S Rahmann, S Nahnsen, and J Köster. Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. *F1000Research*, 10(33), 2021.