Balsam: A Collaborative Platform to Support ML and ML-Ops initiatives

Jakob Engdahl (Statistics Sweden) plus ChatGPT

1. Introduction

The increasing importance of Machine Learning (ML) in everyday operations has necessitated a shift from viewing ML as an innovation activity to treating it as a regular production activity. Addressing this need, the BALSAM project has been initiated with the aim of accelerating this transition in our organization and potentially others.

BALSAM is a collaborative endeavor involving our organization, AI Sweden, and Örebro University. AI Sweden, as the Swedish center for applied AI, brings its expertise in providing guidance and networking within the Swedish AI community. Örebro University, with its explicit focus on AI research, contributes academic rigor and the latest advancements in the field.

The overall goal of the BALSAM project is to facilitate the transition of ML from an innovation activity to a regular production activity. This involves addressing both technical and non-technical challenges, such as culture, skills, ecosystem, data, and organizational capacity for AI.

2. Machine learning beyond innovation initiatives

Scaling machine learning (ML) within an organization is not solely a technical challenge; it's a transformative journey that touches multiple dimensions beyond technology: culture, skills, ecosystem, data, and organization. Each of these dimensions plays a critical role in the successful and sustainable implementation of an ML project.

Culture: Transitioning to a data-driven culture is fundamental for the successful scaling of ML. It involves fostering an environment that encourages curiosity, experimentation, and the embrace of data and ML solutions.

- Skills: The BALSAM platform aims to lower the skills barrier by providing an end-user-ready environment. By handling complex aspects such as package-versioning, security, and versioning, the platform allows a wider group of people to participate in ML initiatives without requiring extensive expertise. This democratization of ML is a key factor in scaling its application across the organization.
- Ecosystem: The broader ecosystem of partners, suppliers, and customers plays a vital role. The federated knowledge base, a key feature of BALSAM, allows leveraging the collective knowledge and experience of the ecosystem, providing resources, expertise, and feedback critical for scaling ML.
- Data: The quality, accessibility, and governance of data are pivotal factors for ML projects. Robust data management practices and reliable data pipelines are prerequisites for the successful application of ML.
- Organization: Organizational structures and processes need to be adaptable to incorporate ML into everyday operations. For example, the governance organization, without needing a

dedicated IT team, could help develop the content of the "Frameworks and Standards" layer, emphasizing the flexibility and adaptability required in this journey.

While these aspects extend beyond the direct capabilities of technology platforms, the BALSAM platform is designed to act as an accelerator. By providing a robust, flexible, and collaborative platform, BALSAM aims to facilitate the broader transformation required to scale ML.

In the following sections, we will detail how the BALSAM platform's architecture addresses these broader challenges while meeting the technical requirements for scaling ML.

3. Similar requirements on statistical production

In the quest to scale machine learning (ML) beyond the realm of innovation, it is crucial to identify the commonalities between the prerequisites for regular statistical production and those of machine learning. Investment in these shared requirements can provide the foundation for expanding the use of machine learning in production environments.

The BALSAM project aims to build a self-service platform that serves dual purposes. Firstly, it enables the training and maintenance of machine learning models. Secondly, it provides a platform for regular processing and analysis of data as part of regular statistical production.

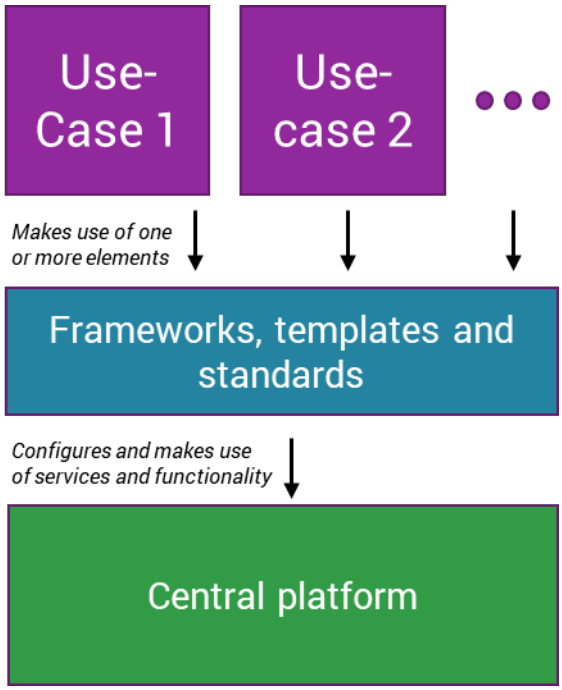
Another important aspect of BALSAM's scope is its support for the evaluation of new data sources. This particular use case shares several requirements with regular production but demands a more flexible platform. Evaluating new data sources can involve web scraping, linking newly sourced data from the web with existing data, and other unique tasks. Therefore, a platform that can accommodate these activities while still maintaining the rigors of statistical production is essential.

However, as we expand the platform's scope, we find that some requirements are universal across use cases, while others vary. For instance, access to the internet and the ability to install R and Python packages are often necessary when evaluating new data sources. However, for regular production, offline execution with limited or no capacity to add packages is the default setting. This approach enhances security and supports a harmonized production environment.

As a guiding principle in the BALSAM project, we allow each project or production round to adapt their settings to fit their specific needs, while still abiding by the overarching policy governing each use-case. This approach ensures that even while projects have the flexibility to tailor their environments, they still maintain the necessary standards and protocols critical for consistency and security. To streamline this process and promote harmonization, we provide templates tailored to different use cases. These templates simplify user adherence to the specific requirements of each project, thereby facilitating both statistical and machine learning production.

4. Layers in the platform

The BALSAM platform's architecture is designed in a layered manner to promote loose coupling between different elements of the system. A high-level overview of this architecture is provided in the following diagram:



Each layer of this architecture serves a distinct purpose and has been crafted to ensure the platform remains flexible, adaptable, and resilient in the face of evolving technological needs.

A critical piece of the BALSAM architecture is its Application Programming Interface (API). This API is the glue that binds the different layers together, allowing for the seamless integration of the 'Frameworks, Templates, and Standards' layer with the central platform. Automation is a core feature of the BALSAM API, aimed at minimizing the need for manual intervention in routine tasks and updates.

This API-centric design allows for significant flexibility in the choice of underlying technologies. Different National Statistical Institutes (NSIs) can have varying infrastructure and service requirements. BALSAM addresses this diversity by making use of open-source services hosted in its GitHub repository, along with the API that provides a mechanism for NSIs to customize the underlying services through the addition of adapters according to their specific needs.

For instance, while BALSAM's GitHub repository includes Minio, an open-source object storage platform, Statistics Sweden has opted to use a different commercial service for Object Storage. To accommodate this variation, BALSAM includes two different adapters in its codebase: one for Minio and another for the commercial platform used by Statistics Sweden. This approach illustrates the platform's commitment to adaptability and customization, enabling it to cater to the diverse needs of different NSIs while maintaining a consistent, user-friendly interface.

4.1 Use-Case level

The Use-case level layer is tailored to accommodate the specific needs and configurations of each project or use-case. While every use-case possesses unique aspects, there exists substantial commonality in their structure and service utilization. This is made possible by leveraging the frameworks, templates, and standards from the underlying layer.

The services employed may vary among typical use-cases, yet Jupyter Notebooks are a ubiquitous feature in the BALSAM ecosystem. These notebooks facilitate multi-language development, making them instrumental in crafting machine learning training code as well as regular data processing and analytical tasks for statistical production. In particular, BALSAM utilizes a version of this platform known as Elyra AI. Elyra AI empowers end-users to sequence a set of functions or scripts, forming an automatable production chain.

BALSAM is designed to support both machine learning projects and statistical production. Statistical production often involves recurring processes with production chains repeated on a regular interval, be it yearly, quarterly, or monthly. To ensure traceability and accountability over multiple survey rounds, we employ versioning and naming standards in the underlying services to track changes in code and data.

Machine learning initiatives are often initially introduced as projects. However, the importance of regular re-training of machine learning models necessitates a repeating structure similar to that of statistical production. In BALSAM, we recognize this need and thus aim to apply the same repeating structure to both machine learning training projects and subsequent re-training rounds. This approach ensures that the machine learning models remain effective and relevant over time.

4.2 Frameworks, templates and standards

This layer, while not fully developed at present, plays an essential role in harmonizing our use of the underlying services and platforms. Furthermore, it expedites the platform's adoption by providing a standard structure to follow. The principal concept of this layer revolves around the provision of pre-configured project templates, which are tailored to suit the different typical use-cases of the platform. Each template comes equipped with a specific set of installed frameworks, unique configurations for network and hardware access, and potentially distinct manuals that guide the usage of the included services and frameworks.

For instance, a few of the project templates we aim to include:

- A Template for Regular Statistical Production: This template could be structured to support the standard processes involved in statistical production, providing the necessary tools and environments for data processing, analysis, and reporting.
- A Template for Machine Learning Projects: This template could be designed with the unique needs of ML projects in mind, providing the necessary tools for data processing, model training, evaluation, and deployment.
- A Template for Innovation and Exploration: This template could be flexible and open-ended, providing a variety of tools and services to support exploratory data analysis, prototyping, and testing of new ideas or methodologies.

By leveraging templates, we can reduce unnecessary variation between projects, ensuring consistent usage of services. This approach also simplifies the user's task of adhering to the relevant policies and standards, leading to more efficient and compliant project execution.

Another important part of this layer is the federated knowledge base. While still under development, this aims at utilizing the power of the community and allows different organisations to contribute with best practice, template projects, code snippets and other technical resources that could promote re-use of existing solutions.

4.3 Central platform

The central platform is a vital component of BALSAM that provides shared functionality across all use-cases and projects, regardless of the specific frameworks, templates, and standards in use. The common functionalities provided by the central platform include:

- Project Management: In BALSAM, all resources are always associated with a specific project. This allows for easy coordination and management of code, services, users, and other resources linked to a particular project.
- Containerization Environment: The central platform provides pre-defined images for creating a containerized environment. This simplifies service creation and ensures that all services run in isolated, reproducible environments.
- Version Control: BALSAM makes use of a Git-based service for versioning of code, documents, and other metadata. Currently, we use GitLab for this purpose.
- Data Versioning: We implement data versioning through object storage functionality. We currently use an S3 compatible service called Minio for this. However, we also plan to introduce a commercial object storage service for our internal use. Despite this, we intend to continue releasing a version with Minio to support users who prefer this open-source solution.
- Network Accessibility Policies: The central platform allows for the implementation of different network accessibility policies. Depending on the template, code execution may be entirely offline, have access to specific internal services, or have access to the internet.
- Hardware Scaling Performance: Different templates cater to varying hardware performance requirements. Some templates are designed for projects that require more memory and GPUs, while others are suitable for projects requiring only basic hardware capacity. We are currently establishing a new data center with higher GPU capacity. This enhanced functionality will be accessible once we switch to the new data center.
- User Authorization and Authentication: BALSAM employs role-based security, with a limited number of roles available. The roles a user has may vary across projects, aligning with the specific needs and responsibilities within each project.
- Collaboration platforms for communication between roles and organizations within a project. We also aim at integrating the processing environment with the collaboration platform to enable updates from production execution within the project communication channels.

5. Current status and further development

The BALSAM project has been underway for approximately one and a half years now. Our current goal is to have a Minimal Viable Product (MVP) ready by the end of the year. Much of our focus has been on developing the central platform, as it incorporates new technological frameworks that are novel to Statistics Sweden.

In order to ensure that the platform meets the needs of various use-cases, we have launched several pilot projects. These pilots utilize the platform in its current state and provide invaluable feedback to guide our ongoing development efforts.

The middle layer of the platform, concerning the standardization of frameworks, templates, and standards, still requires more work. This layer is critical as it sets new requirements on how the central platform exposes its functionality. The intent is to create a more harmonized and user-friendly environment that facilitates quicker adoption of the platform across different use-cases.

As we progress with the BALSAM project, our aim is to continually improve the platform and expand its capabilities to better serve the needs of Statistics Sweden and in extension, other NSIs and ONAs.