



The Role of Open Source for Data Science, Engineering and Programming Teams in the Pharma Industry

DH09 – Tadeusz Lewandowski & Axel Muehlig, Data Engineering team at Janssen Research & Development



Agenda

- What is open source
- What are the opportunities?
- What is a role of open source in Pharma?

What is open source

Skin cells at 20x magnification

What is open source

Definition of open source

Open source definition is more comprehensively captured and includes following criteria, and does not simply mean "*access to the source code*"¹:

- Free Redistribution – no restriction on distribution
- Source Code – must include the source code
- Derived Works – must allow modifications
- Integrity of The Author's Source Code - must explicitly permit distribution
- License Must Not Restrict Other Software
- Licenses types by category: <https://opensource.org/licenses/category>

1 - [The Open Source Definition](#)

What is open source

What open source offers

- Open source offers various products: [Linux](#), [R](#), [Python](#), [Apache Spark](#) and [Hadoop](#), [CI/CD tools](#) or [Git](#)
- Open source is triggering democratization of accessing, utilizing and updating solution/projects by individuals, small teams or collaborations via established communities:
 - [R Consortium](#)
 - [Bioconductor](#)
 - [Neuroconductor](#)
 - [ROpenSci](#)
 - [openpharma](#)

What is open source

How big is the open source?

- GitHub – is popular host open source software development projects.
- As of January 2023, GitHub reported having²:
 - **over 100 million developers and**
 - **372 million repositories,**
 - **at least 28 million public repositories**

GitHub license recommendation: *“Public repositories on GitHub are often used to share open source software. For your repository to truly be open source, you'll need to license it so that others are free to use, change, and distribute the software”*³

What are the opportunities?

Skin cells at 20x magnification

What are the opportunities?

What are the different categories

Open source provides:

Tools/Solutions: a range of ready to use or flexible solutions, also in combinations with commercial tools

Documentation: a significant part of the open source code are: great documentation and examples

Webinars: Tutorials and free online courses complements it and enables a quick implementation

Communities: Have important role in contributing, sharing ideas and enabling influencing existing solutions e.g.: pharmaverse.org

What are the opportunities?

Xops – managing workflows

What are the opportunities for statistical programmers, data engineers and data scientists

Open source is multidisciplinary as is: combining methodologies, tools and technology organisation

- DevOps⁴ - *methodology introduced by software development bringing operations and development of software to be deliver more efficiently*
- DataOps⁴ - *Data engineer with supporting tools, performing data transformation activities and building data workflows*
- MLOps⁴ - *simplification of management and deployment of machine learning models*

What is a role of open source in Pharma?

Skin cells at 20x magnification

What is a role of open source in Pharma?

DevOps – managing workflows

DevOps⁵ - methodology introduced by software development bringing operations and development of software to be deliver more efficiently

Clinical and Statistical programmers can leverage good practices and operations like:

- Using the CI/CD framework:
 - Improved code source control
 - Building automation tools on programming dependencies
 - Testing automatically code base
- Employ Agile methodology in daily work practice:
 - Stand-ups, sprints, retrospective

What is a role of open source in Pharma?

From DevOps to DataOps

DevOps emerged over time as traditional development overcoming complexity of software development and dependencies, the same principles could be applied to DataOps

Implementing a new set of practices and tools to scale-up data engineering by implementing:

- PEOPLE: Agile methodology to improve team's collaboration and build in continues improvements
- PROCESS/TOOLS: selecting “fit-for-purpose” tools and methods and deliver data operation increments
- TECHNICAL: Automation in data deployment leading to data quality and overall data integration

What is a role of open source in Pharma?

DataOps – managing data workflows

Delivering reliable data continuously, and extracting new knowledge via data-driven techniques

- **DataOps** enables:
 - Faster development iterations
 - Automating repeated data operations
 - Improvements data quality
- **DataOps** could be enabled by following steps⁶
 - Setting new governance model for different DataOps steps, mandating automation and integration with the business needs
 - Building highly-skilled data engineers teams
 - Building tailored tools in order to support data operations.

What is a role of open source in Pharma?

DataOps – workflows

- DataOps in open source offers variety of tools⁶:
Cataloguing, exploration, ingestions, workflows, Data processing, Data quality and visualizations
- The DataOps can be summarized by the DataOps manifesto⁷: *Continually satisfy your customer, Embrace change, Daily interactions, self-organize, Reflect, Orchestrate, Make it reproducible, Reuse, Improve cycle times ...*

Modern DataOps at enterprise level might require significant infrastructure changes and amendments in order to response to growing needs in establishing data-driven organisation and leveraging data assets effectively.

6: [Example of curated DataOps tools](#), 7 – [DataOps manifesto](#)

What is a role of open source in Pharma?

MLops – managing workflows

- MLOps - simplification of management and deployment of machine learning models
 - Unifies release cycles for ML
 - Enables automated testing
 - Models' deployment

Machine Learning engineering using MLOps adds the value to DataOps and focusing on streamlining process of taking machine learning models to production

Conclusions

Skin cells at 20x magnification

Conclusions

Potential advantages

- Cost-effectiveness by utilizing open source solutions
- Collaboration and flexibility in the development new solutions
- Attracting talents to data-driven Pharmaceutical industry
- Faster decision making, based on the complex clinical data