



Single Day Event

Arwa Topiwalla
Saama Technologies



Regulatory Data Submissions: Cost Effective using open Source Technology

When we think about creating clinical trial regulatory submission datasets (SDTM/ADaM) we often only think of closed sourced, expensive and licensed software. A majority of the regulatory submissions are done using these licensed software, however there are other alternatives for clinical and bio-statistical programming.

The objective of this presentation is to bring out alternatives to this traditional approach by discussing open source technologies which provides all the functionality required to create CDISC compliant datasets and cater to regulatory submissions.

Here we will discuss: Are expensive technologies required for creating a CDISC submission package and what are the advantages of open source technologies for related/common business use cases.



Regulatory Data Submissions: Cost Effective using open Source Technology

When it comes to CDISC SDTM programming, 99% of programmers use SAS and identify themselves as SAS programmers.

Most of the Pharmaceutical companies use SAS and the CROs/Business providers continue this trend. The main reason for this is

- Inbuilt features that aid the programmers to output CDISC and regulatory compliant data.
- Thoroughly tested and stable software
- Support from SAS
- SMEs are proficient in SAS due to which the trend follows.

Open Source Technology Advantages



Business Use Case

1. Cost

Open source is available for free. This is in stark contrast to costly licensed software in a closed environment, where programmers cannot use it without a proper license.

2. Accessibility

R and Python have been on the frontlines with a lot of packages/libraries related to machine learning. The main reason is that anyone can operate on them as it is open source.

3. Connectivity

Licensed software often requires users to connect to a separate server. They are not usually available on our desktops for quick usage due to cost concerns.



Functionality

1. Libraries & Community

Python has a plethora of packages that developers can use to add features in the any application.

e.g Numpy or Pandas (Python), R - data handling.

Scipy (Python) - scientific formulas

PHP - API development

2. AI/ML

Open source technology like Python is the largest community for AI ML and has package support for building ML models.

e.g Tensorflow and *Keras(C, C++, Python)* - Huge and stable libraries helping python for deep learning

Sci-kitlearn(Python) - Machine Learning



Functionality

3. Scalable

The biggest benefit of open source technologies like Python is that it is a highly scalable programming language.

e.g Spark(Scala, Python) - Handling large data

4. Readability

Python is one of the programming languages used in healthcare that has a clear and easily readable syntax. Even for adding complex functionalities, developers can write simple code or add plugins for making it feature-rich.



Product Building

1. Development

Developing programming API with various options is possible in open source programming.

Microservices built in multiple programming languages can communicate with each other using open source languages. This makes sure the product is modular and has the best features.

*e.g. **Flask**, **DJango (Python)**, **PHP**, **Nodejs**, **Express***

2. Testing

Testing packages are available that helps to develop test cases in such a way that product testing hindrances can be reduced.

*e.g **Pytest(Python)** - Product testing*

***Unittestcase(Python)** - Unit Testing*



Product Building

3. Automation

Open Source Technologies like Python can be used for lot of Automation in the product which deals with excel sheets, codelist, standards, pinnacle 21 errors.

4. Job

There are many packages available for Job automation.

e.g. **Apscheduler(Python)** - Scheduling jobs

5. Program builder

Python and R packages can be used to read mapping spreadsheets and build the program (in Python, R or SAS)

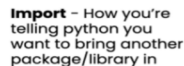


Open Source Technology for Submission

- SDTM Specifications
e.g. ***SDTM-Mapper(Python)***
- CDISC datasets Programming
e.g. ***Pandas, Pyucum (Python), Admiral(R)***
- XPT creation
e.g. ***XPORT(Python)***
- Define Excel
e.g. ***Pandas(Python)***
- ACRF annotations
e.g. ***PDF-Annotate***
- Visualizations can be created easily using various packages
e.g. ***Plotly, Seaborn, R-Studio***
- Edit checks
e.g. ***Pandas(Python)***



- Over 1000+ optimized implementation of functions, readily available with one import command.
- Dataframe structure in pandas for tabular data wrangling.
- Dynamic allocation of data-types based on source data.
- Implementation of merge for different data scenarios with low time complexity.
- Ease Date and Time modification with Datetimeindex and Timedeltaindex functions.



As - Creating an alias. 'As' tells python that you're going to call "pandas" as "pd" from now on

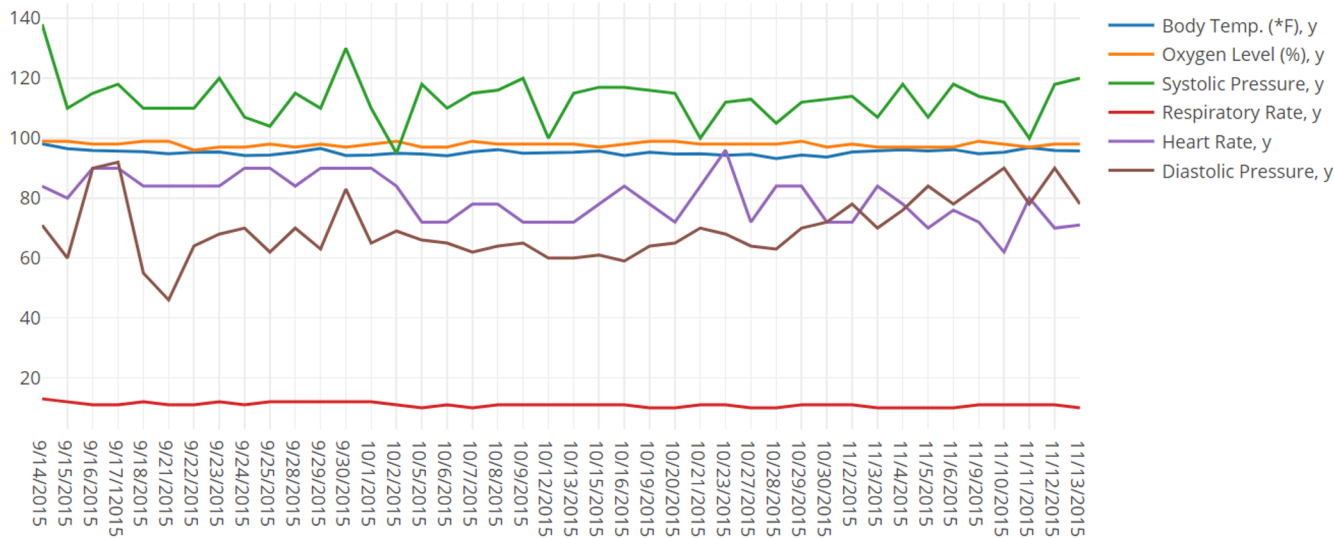
pandas - The package/library you want to import. In this case, it's pandas

pd - The standard way to reference pandas. This means instead of typing "pandas.function" you'll only need to write "pd.function"



Open Source Technology for Submission

Visualization in Plotly for Data Classification





Open Source: Eco-System

At Saama, we are leveraging open source technologies to provide our sponsors with flexible and cost-effective solutions, without sacrificing on advanced features and for creating future products



kubernetes

Production-Grade Container Orchestration



Programming Language



Jenkins

Build great things at any scale



Apache
Airflow

Job Management



Open Source: Eco-System

Jenkins - Is a leading open source automation server, it provides hundreds of plugins to support building, deploying and automating any project.

Advantage: Continuous Integration and Continuous Delivery

As an extensible automation server, Jenkins can be used as a simple CI server or turned into the continuous delivery hub for any project.

Kubernetes (K8) - is used for automating deployment, scaling, and management of containerized applications.

Advantage : Allows Parallel processing, faster execution lesser wait time. Faster system deployment. System configuration control and Secure environment.



Open Source: Eco-System

Apache Airflow - is a platform created by the community to programmatically author, schedule and monitor workflows.

Advantage: Keeps track of how many mapping jobs and submission jobs are running and status of the jobs

Apache Airflow does not limit the scope of your pipelines; you can use it to build ML models, transfer data, manage your infrastructure, and more.

Pyspark - It is majorly used for processing structured and semi-structured datasets.

Advantage: Large data handling and lesser wait time for execution. It also provides an optimized API that can read the data from the various data source containing different files formats

Thank-You



Arwa Topiwalla
Associate Director - Saama Technologies

arwa.topiwalla@saama.com

15 + years of experience in Life Sciences as a Biostatistical programmer. Proficient in statistical programming activities for phase I-IV clinical trials including CDSIC and submission packages. Part of the standards team to set up global standards for CDSIC and TFLs. Currently leading the standards team for Saama's smart series SAM (Smart Auto mapper) for CDISC (SDTM+ADaM) and TFLs



**The Global Healthcare
Data Science Community**

Contact Channels

UK +44 1843 609600
office@phuse.global
phuse.global

Social Media

@phusetwitta
/phusebook
/phusetube
/company/phuse