Dataset-JSON Viewer

When Machine-Readable Becomes Human-Viewable

Dmitry Kolosov

14 May 2025

Agenda

Motivation

- What am I doing wrong
- COSA Hackathon

Viewer

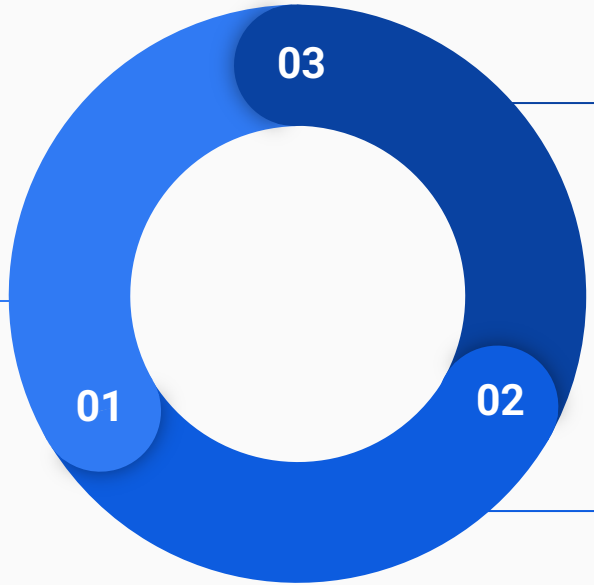
- Functionality Overview

Next-Generation of Viewers

- Data View
- Data Validation
- Data Analysis

Motivation

Development Process

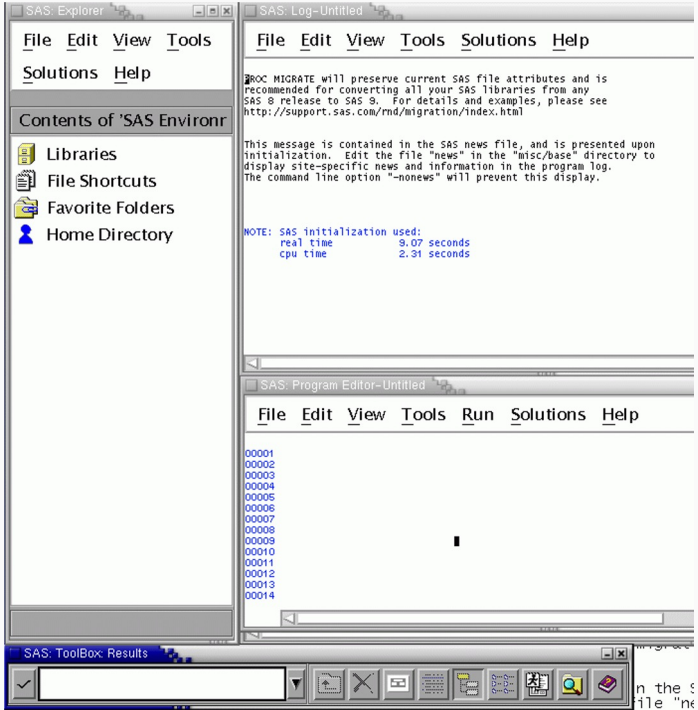


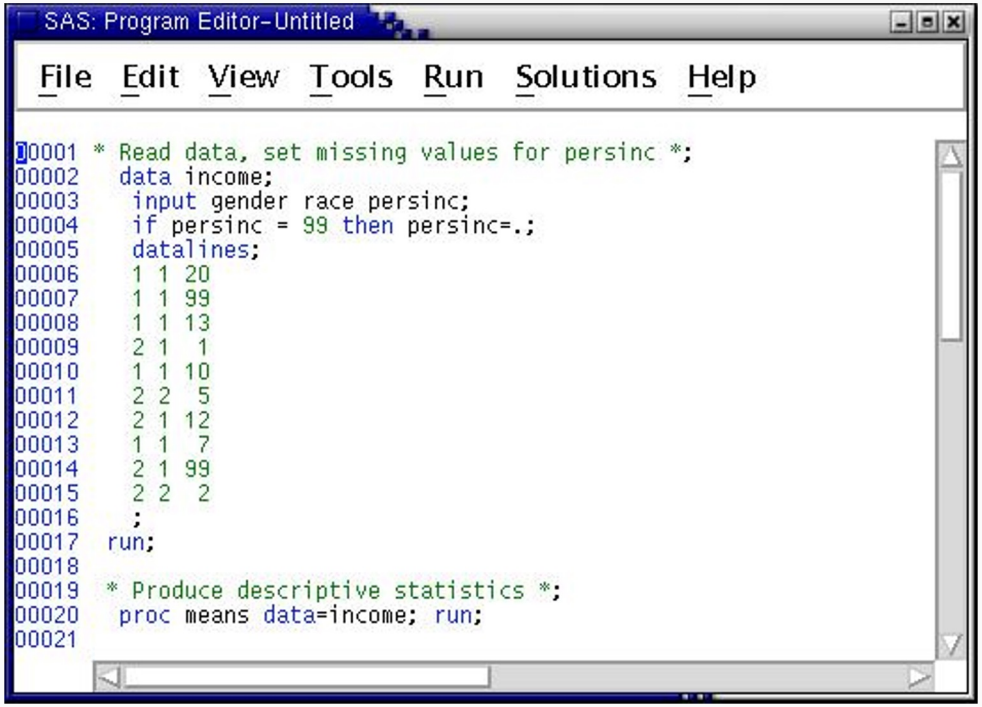
Develop Code
IDEs, Editors, Macro Libraries, Packages

Review
Dataset Viewers, Validators

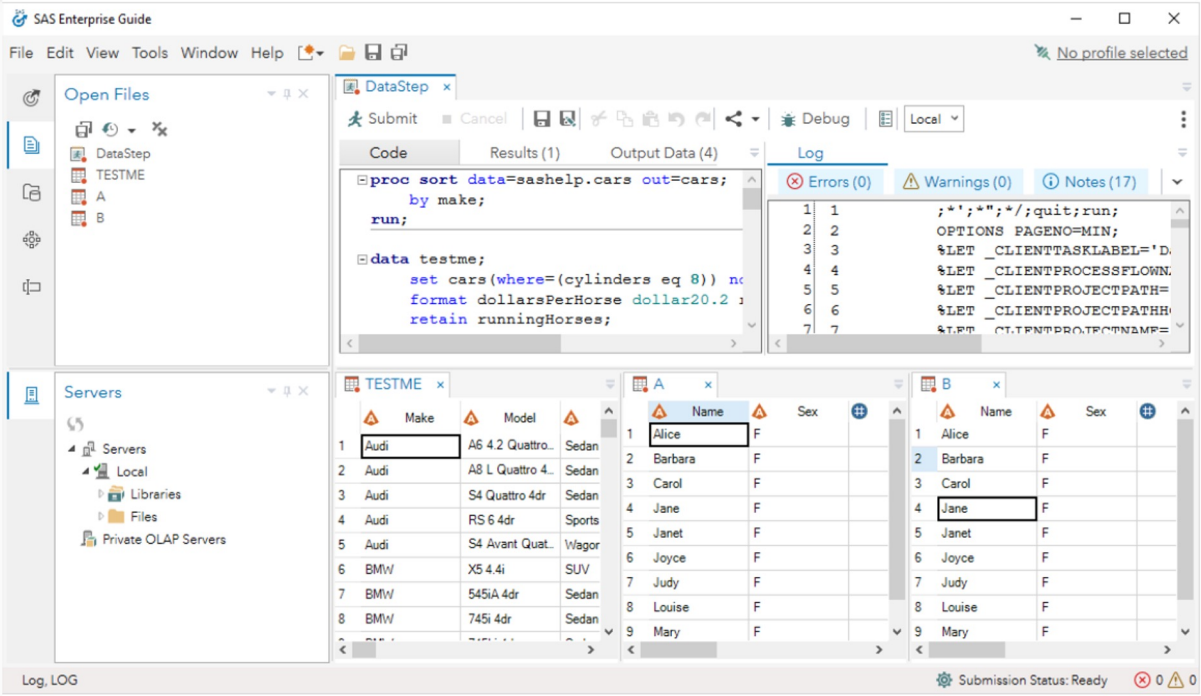
Execute
Servers, Languages

Development Process

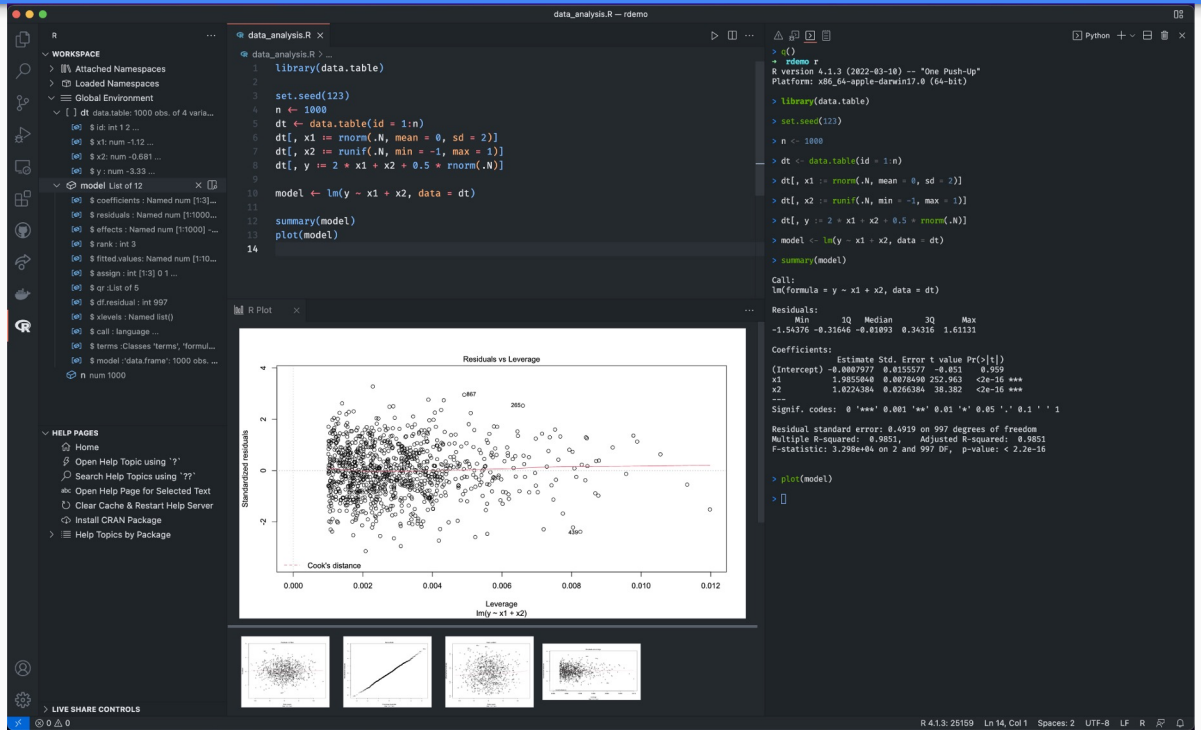




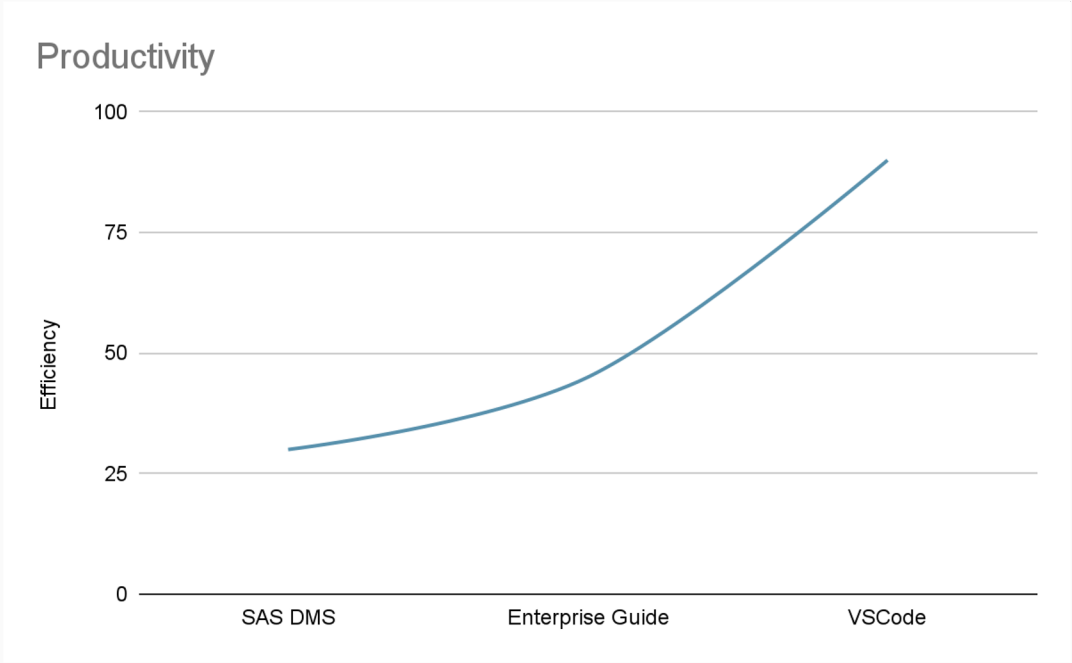
Development Process



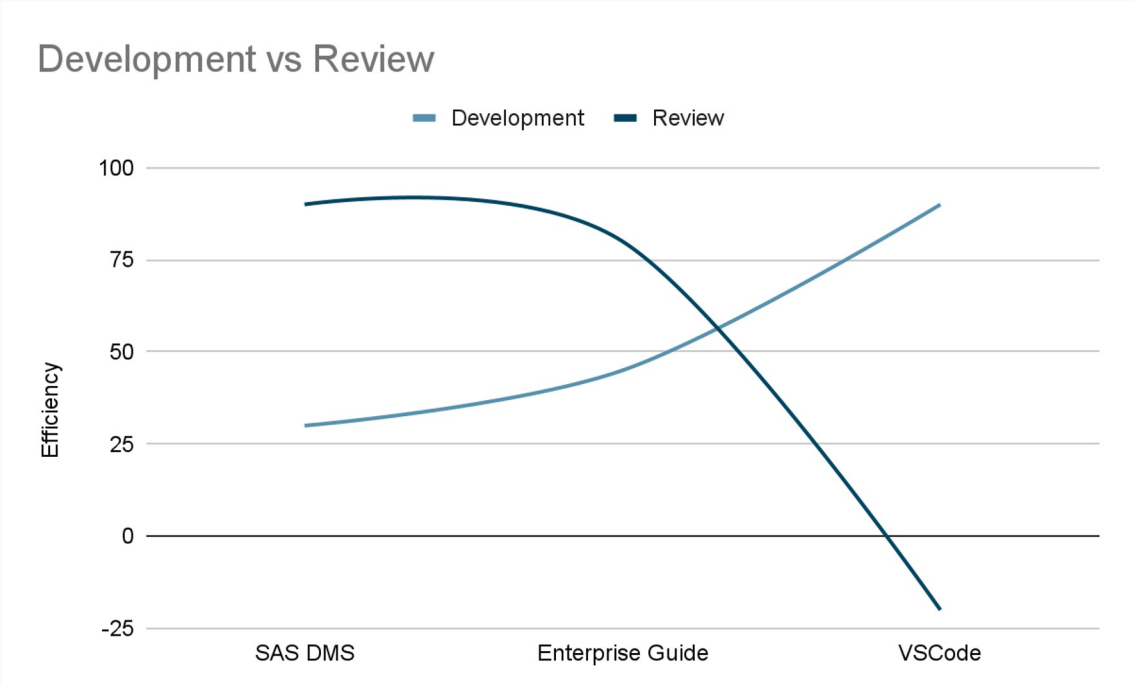
Development Process



Development Process



Development Process



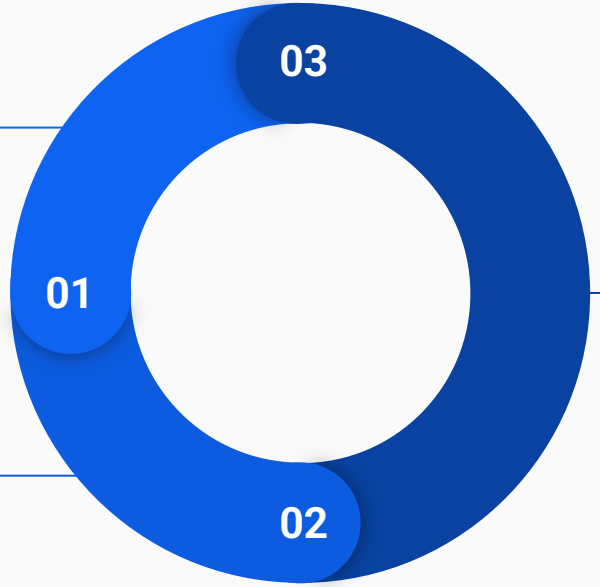
Development Process

Develop

IDEs, Editors, Macro Libraries, Packages

Execute

Servers, Languages



Review

Dataset Viewers, Validators

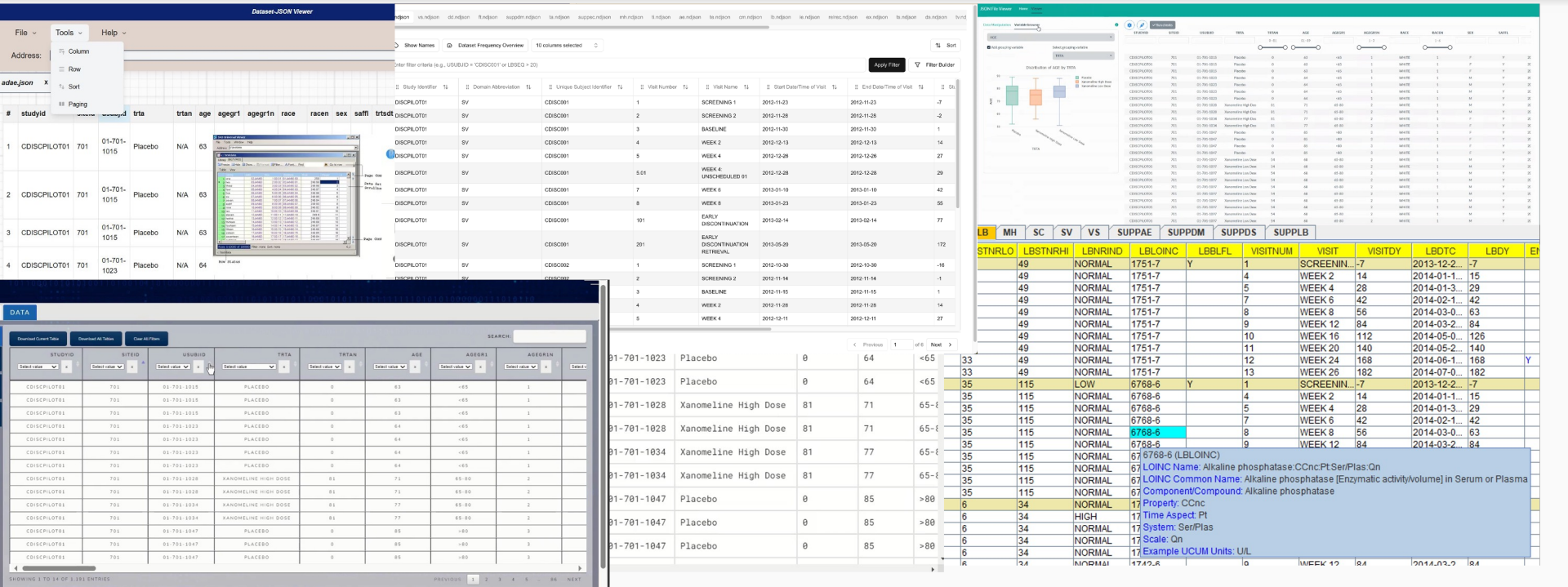
COSA Hackathon

- **Primary objective**: Create Dataset-JSON Viewer software

3 month of development

7 submitted projects

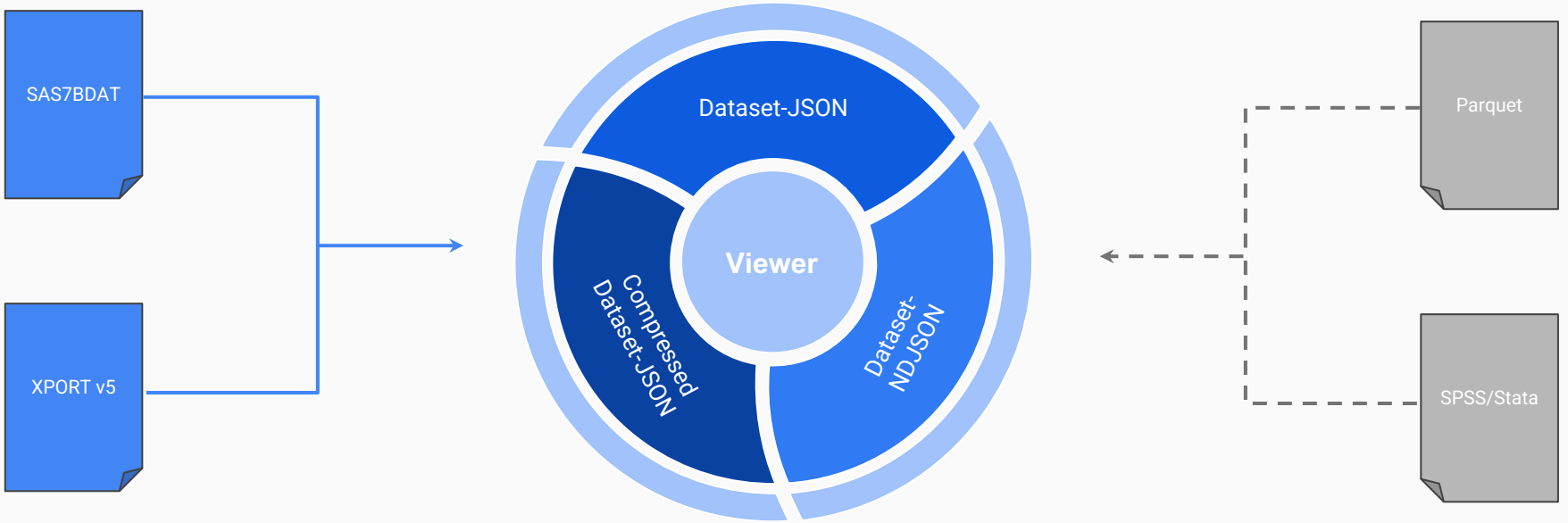
COSA Hackathon



Desperate Programmers

VDE Dataset Viewer

Data Sources



Viewer

- Desktop App
  - Windows
  - Linux
  - MacOS (can be compiled)
- Filters
  - Manual/Interactive filters
  - Multiple operator and functions
  - Autocompletion
  - Validation
  - Last 100 filters
- Navigation
  - Go To Column/Row/Column and Row
  - Virtualization (load millions or rows and thousands or columns)
  - Switching between datasets
- Metadata View
  - Dataset Metadata
  - Variable Metadata
  - Unique Values and Counts
- Streaming
  - Load infinitely large datasets
  - Quick open
- Miscellaneous
  - Selection copy
  - Quick filters
  - Column Masks
  - Numeric Value Rounding

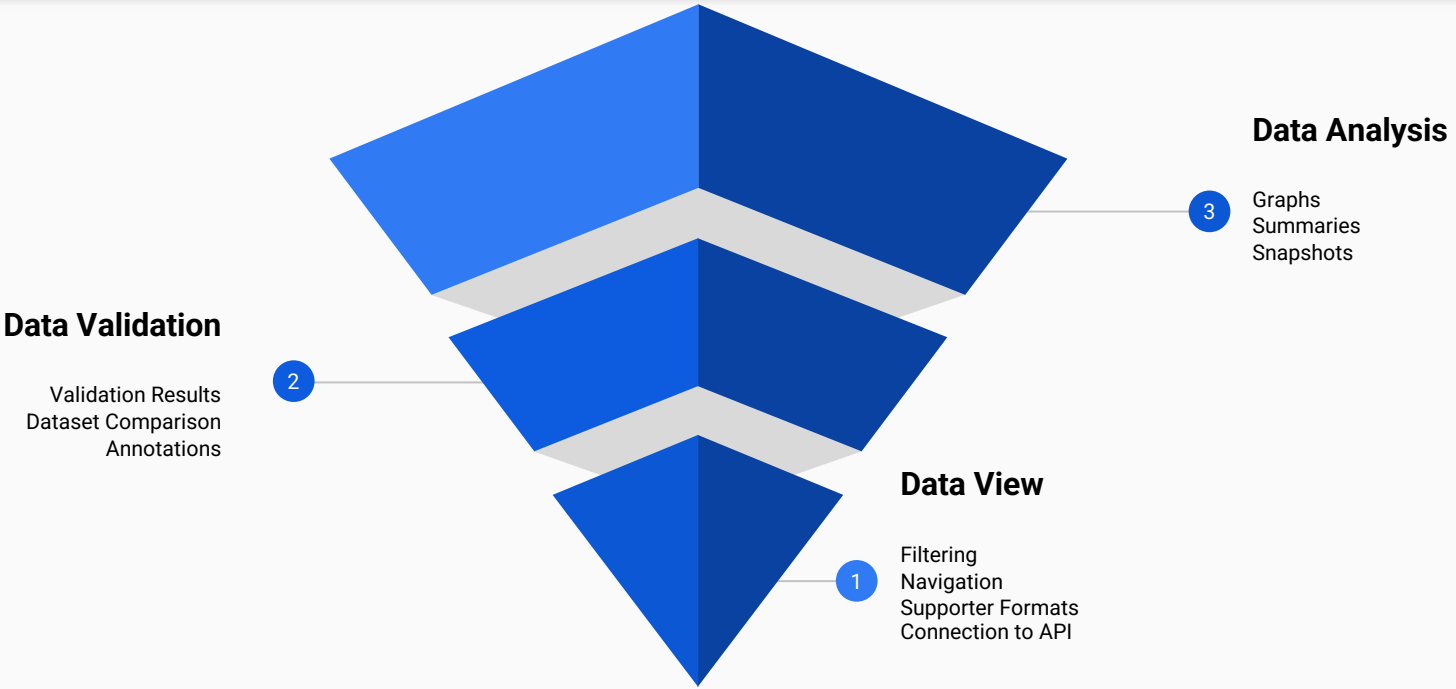
Additional Functionality

- Supported Data Formats
  - Dataset-JSON 1.1
  - Dataset-NDJSON 1.1
  - Compressed Dataset-JSON 1.1 (Prototype)
  - XPORT v5
  - SAS7BDAT
- Dataset-JSON API

- Converter
  - Bulk processing
  - Multiple formats support
  - Setting Metadata
  - Variable type detection

Next Steps

Next Generation Data Viewer



More Formats and Data Sources

Support to additional formats and ways to load data

Further develop Dataset-JSON API to support all CRUD operations

Currently Supported formats and sources

- Dataset-JSON 1.1
- Dataset-NDJSON 1.1
- Compressed Dataset-JSON 1.1 (Prototype)
- XPORT v5
- Dataset-JSON API server
- SAS7BDAT

Chess

- Millions of Users
- Multiple Platforms
- Lots of features

Chess

Chess.com

Lichess.org

- 650+ employees
- Team is split over 60 counties
- Revenue over $100 million

Data Validation

Dataset Comparison

Integration with Data Validation tool from CDISC

Dataset Comparison

Essential part of double programming is dataset comparison

- Differences summary
- Side by side view with differences highlighted

Data Validation

CDISC CORE project has UI part which allows to execute checks in a browser

- Run check for the current dataset
- View Issues in Dataset
- Annotate Issues

Data Analysis

Analyzing Data

Finding an issue in dataset is similar to finding a needle in a haystack

Many issues are identified by:

- Double Programming
- Data Validation
- Summary outputs review

Multiple open-source tools are developed as part of Pharmaverse to support data analysis and visualization.

- Fast generation
- Traceability back to the source data
- R (WebR)/Python

The End

defineEditor.com

https://github.com/defineEditor/vde-dataset-viewer