

Dataset-JSON Viewer

When Machine-Readable Becomes Human-Viewable

Dmitry Kolosov

14 May 2025



Agenda

Motivation

- What am I doing wrong
- COSA Hackathon

Viewer

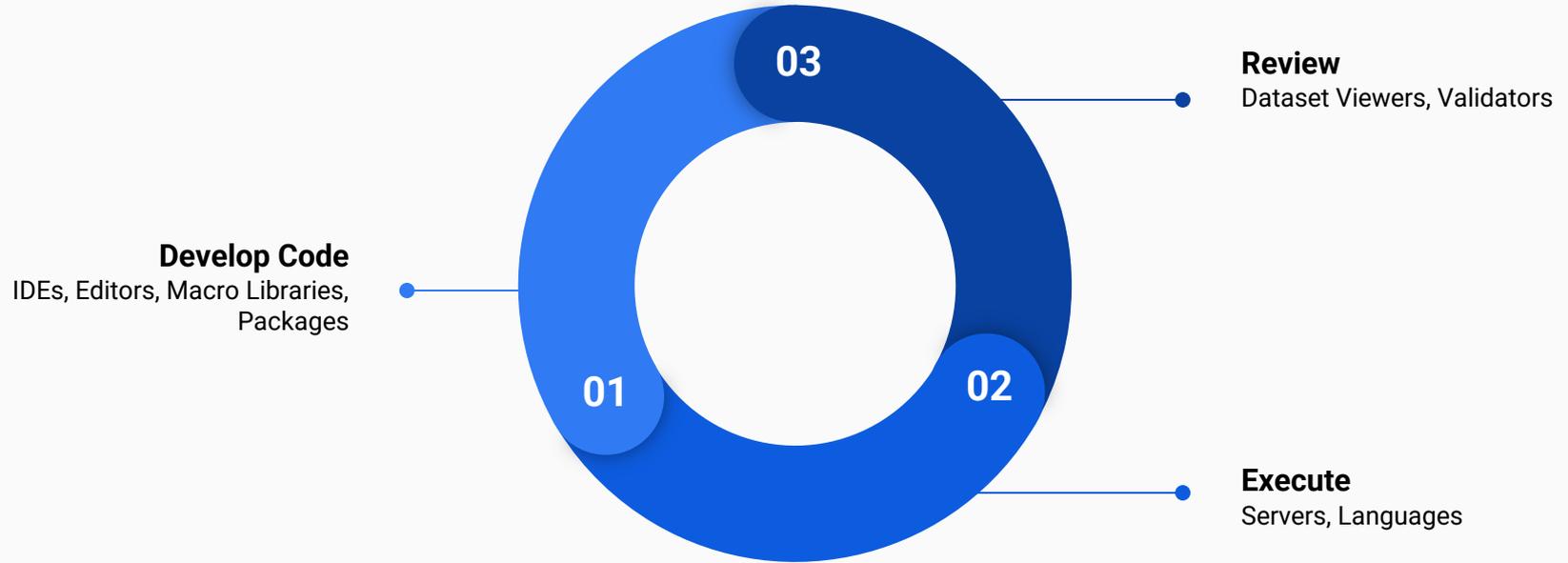
- Functionality Overview

Next-Generation of Viewers

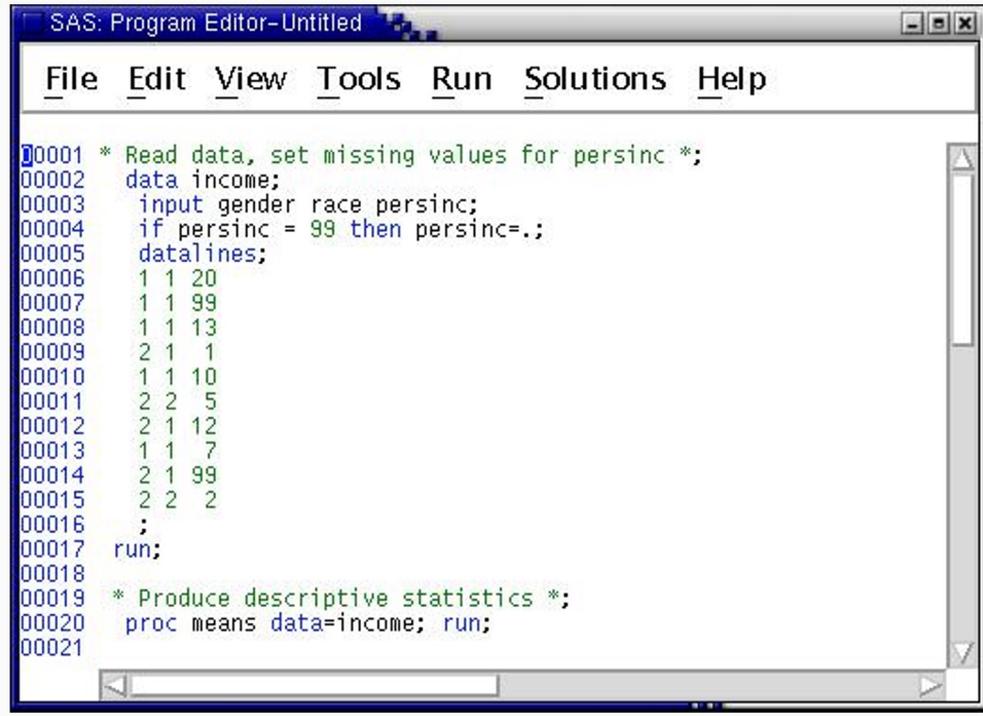
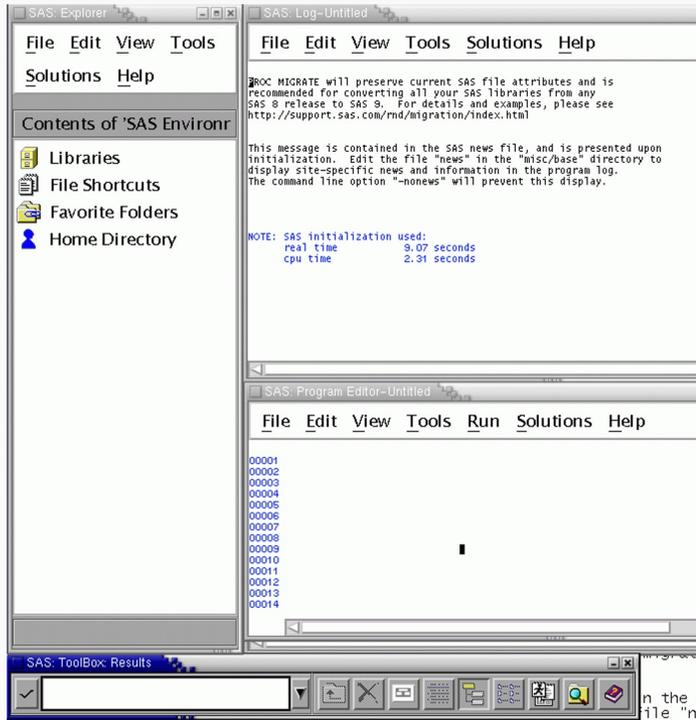
- Data View
- Data Validation
- Data Analysis

Motivation

Development Process



Development Process



Development Process

The screenshot displays the SAS Enterprise Guide interface. The top menu bar includes File, Edit, View, Tools, Window, and Help. The main workspace is divided into several panes:

- Open Files:** Shows a list of files including DataStep, TESTME, A, and B.
- Servers:** Shows a tree view of servers, including Local, Libraries, and Private OLAP Servers.
- Code Editor:** Contains the following SAS code:

```
proc sort data=sashelp.cars out=cars;
  by make;
run;

data testme;
  set cars(where=(cylinders eq 8)) nobs;
  format dollarsPerHorse dollar20.2;
  retain runningHorses;
```
- Log:** Shows the execution log with the following content:

```
1 1 ;*;*;*;/quit;run;
2 2 OPTIONS PAGENO=MIN;
3 3 %LET _CLIENTTASKLABEL='D';
4 4 %LET _CLIENTPROCESSFLOWNUM=1;
5 5 %LET _CLIENTPROJECTPATH=
6 6 %LET _CLIENTPROJECTPATHH=
7 7 %LET _CLIENTPROJECTNAME=
```
- Data Tables:** Three data tables are displayed below the code editor:
 - TESTME:** A table with columns Make, Model, and Sex. The first row is highlighted.
 - A:** A table with columns Name and Sex. The first row is highlighted.
 - B:** A table with columns Name and Sex. The fourth row is highlighted.

The bottom status bar shows "Log, LOG" and "Submission Status: Ready".

Development Process

The screenshot displays the RStudio interface with the following components:

- Workspace:** Shows the loaded environment with a data table of 1000 observations and a linear model object.
- Code Editor:** Contains the following R script:

```
1 library(data.table)
2
3 set.seed(123)
4 n <- 1000
5 dt <- data.table(id = 1:n)
6 dt[, x1 := rnorm(.N, mean = 0, sd = 2)]
7 dt[, x2 := runif(.N, min = -1, max = 1)]
8 dt[, y := 2 * x1 + x2 + 0.5 * rnorm(.N)]
9
10 model <- lm(y ~ x1 + x2, data = dt)
11
12 summary(model)
13 plot(model)
14
```
- Console:** Shows the execution output, including the R version (4.1.3), platform (x86_64-apple-darwin17.0), and the summary of the linear model:

```
Call:
lm(formula = y ~ x1 + x2, data = dt)

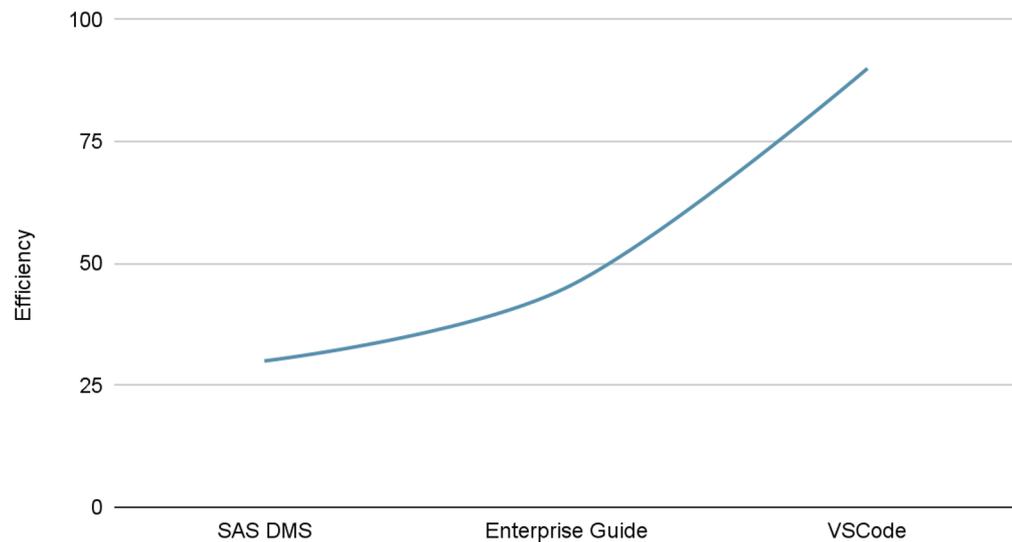
Residuals:
    Min       1Q   Median       3Q      Max
-1.54376  -0.31646  -0.01893   0.34316   1.61131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0807277  0.0355577  -0.851   0.959
x1           1.2855848  0.0076498  152.563 <2e-16 ***
x2           1.0224384  0.0266384   38.382 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4919 on 997 degrees of freedom
Multiple R-squared:  0.9851, Adjusted R-squared:  0.9851
F-statistic: 3.298e+04 on 2 and 997 DF, p-value: < 2.2e-16
```
- Plots:** A central plot titled "Residuals vs Leverage" shows standardized residuals on the y-axis and leverage (lm(y ~ x1 + x2)) on the x-axis. Two points are highlighted with labels: 287 and 280. Below the main plot are four smaller diagnostic plots: a scatter plot of y vs x1, a scatter plot of y vs x2, a Q-Q plot of residuals, and a Cook's distance plot.

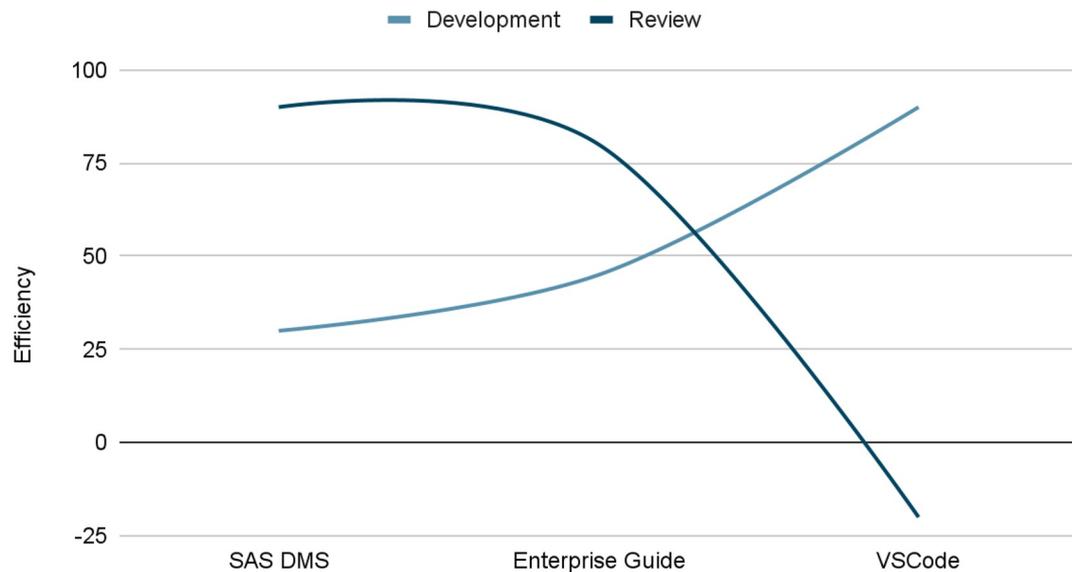
Development Process

Productivity

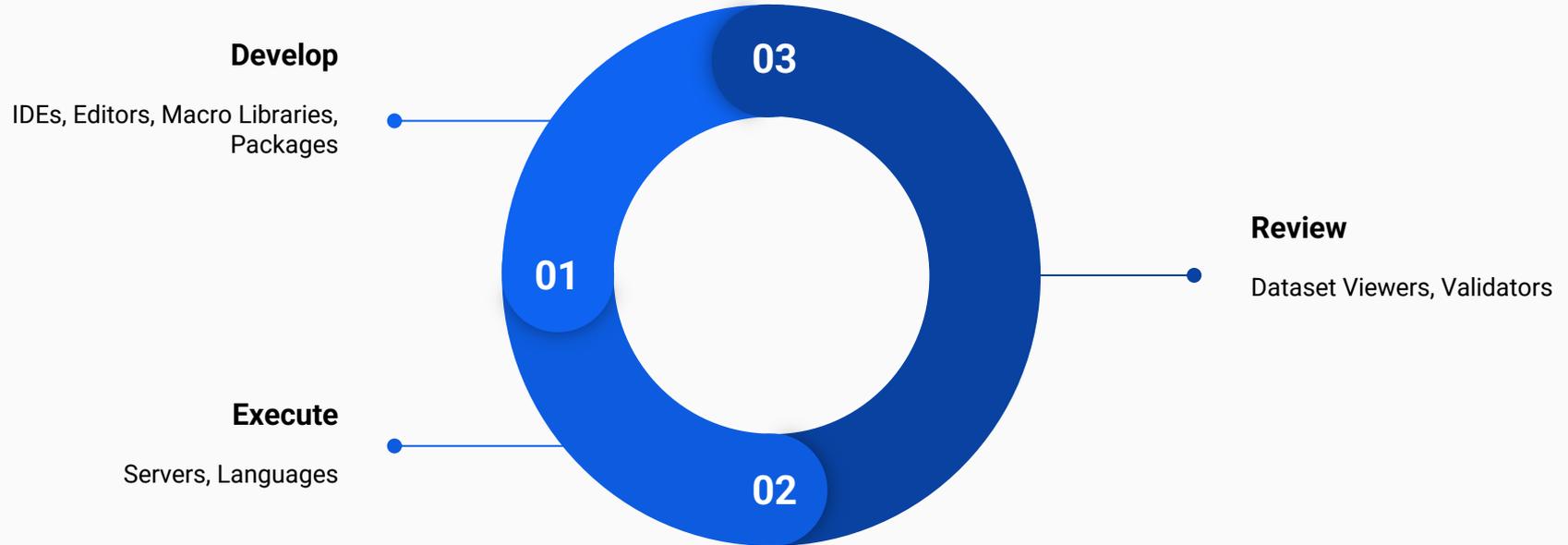


Development Process

Development vs Review



Development Process



COSA Hackathon

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



3 month of development

7 submitted projects



COSA Hackathon

Dataset-JSON Viewer

Tools: Column, Row, Filter, Sort, Page

Address:

Filter: Filter on fields (e.g., USUBJID = 'CDISC001' or LBSEQ = 20)

Apply Filter

#	StudyID	TRTA	TRTAN	AGE	AGEGR1	AGEGR1N	RACE	RACEN	SEX	SFFM	TRSDT	DISCPLT01	SV	CDISC001	E_Unique Subject Identifier	E_Visit Number	E_Visit Name	E_Start Date/Time of Visit	E_End Date/Time of Visit	E_Skipped
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	1	SCREENING 1	2012-11-23	2012-11-23	-7	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	2	SCREENING 2	2012-11-28	2012-11-28	-2	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	3	BASELINE	2012-11-30	2012-11-30	1	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	4	WEEK 2	2012-12-13	2012-12-13	14	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	5	WEEK 4	2012-12-28	2012-12-28	27	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	5.01	WEEK 4 UNSCHEDULED 01	2012-12-28	2012-12-28	29	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	7	WEEK 6	2013-01-10	2013-01-10	42	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	8	WEEK 8	2013-01-23	2013-01-23	56	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	101	EARLY DISCONTINUATION	2013-02-14	2013-02-14	77	
1	CDISCPILOT01	701	01-701-1015	Placebo	N/A	63						DISCPLT01	SV	CDISC001	201	EARLY DISCONTINUATION RETRIEVAL	2013-05-20	2013-05-20	172	
4	CDISCPILOT01	701	01-701-1023	Placebo	N/A	64						DISCPLT01	SV	CDISC002	1	SCREENING 1	2012-10-30	2012-10-30	-16	
4	CDISCPILOT01	701	01-701-1023	Placebo	N/A	64						DISCPLT01	SV	CDISC002	2	SCREENING 2	2012-11-14	2012-11-14	-1	
4	CDISCPILOT01	701	01-701-1023	Placebo	N/A	64						DISCPLT01	SV	CDISC002	3	BASELINE	2012-11-16	2012-11-16	1	
4	CDISCPILOT01	701	01-701-1023	Placebo	N/A	64						DISCPLT01	SV	CDISC002	4	WEEK 2	2012-11-28	2012-11-28	14	
4	CDISCPILOT01	701	01-701-1023	Placebo	N/A	64						DISCPLT01	SV	CDISC002	5	WEEK 4	2012-12-11	2012-12-11	27	

DATA

SEARCH

STUDYID	USUBJID	TRTA	TRTAN	AGE	AGEGR1	AGEGR1N	ALBINC
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	71	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	71	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	71	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	77	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	77	65-80	2
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3

LB	MH	SC	SV	VS	SUPPAAE	SUPPDM	SUPPDS	SUPPLB		
STNRLO	LBSTNRHI	LBHRIND	LBLOINC	LBBLFL	VISITNUM	VISIT	VISITDY	LBOTC	LBODY	ET
49	NORMAL	1751-7	Y	1	SCREENING	-7	2013-12-2	-7		
49	NORMAL	1751-7		4	WEEK 2	14	2014-01-1	15		
49	NORMAL	1751-7		5	WEEK 4	28	2014-01-3	29		
49	NORMAL	1751-7		7	WEEK 6	42	2014-02-1	42		
49	NORMAL	1751-7		8	WEEK 8	56	2014-03-0	63		
49	NORMAL	1751-7		9	WEEK 12	84	2014-03-2	84		
49	NORMAL	1751-7		10	WEEK 16	112	2014-05-0	126		
49	NORMAL	1751-7		11	WEEK 20	140	2014-05-2	140		
49	NORMAL	1751-7		12	WEEK 24	168	2014-06-1	168	Y	
49	NORMAL	1751-7		13	WEEK 26	182	2014-07-0	182		
35	115	LOW	6788-6	Y	1	SCREENING	-7	2013-12-2	-7	
35	115	NORMAL	6788-6		4	WEEK 2	14	2014-01-1	15	
35	115	NORMAL	6788-6		5	WEEK 4	28	2014-01-3	29	
35	115	NORMAL	6788-6		7	WEEK 6	42	2014-02-1	42	
35	115	NORMAL	6788-6		8	WEEK 8	56	2014-03-0	63	
35	115	NORMAL	6788-6		9	WEEK 12	84	2014-03-2	84	
35	115	NORMAL	6788-6 (LBLOINC)		17	Property: CCnc				
35	115	NORMAL	6788-6 (LBLOINC Name: Alkaline phosphatase CConcPtSerPlas) On		17	Time Aspect: Pt				
35	115	NORMAL	6788-6 (LBLOINC Common Name: Alkaline phosphatase [Enzymatic activity/volume] in Serum or Plasma		17	System: SerPlas				
35	115	NORMAL	6788-6 (Component/Compound: Alkaline phosphatase		17	Scale: On				
6	34	NORMAL	17	Example UCUM Units: U/L						

ADAPT to Viewer

Variables: Variable browser, Distribution of Adt by TRTA, TRTA

Apply Filter

STUDYID	USUBJID	TRTA	TRTAN	AGE	AGEGR1	AGEGR1N	ALBINC
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1015	PLACEBO	0	63	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	PLACEBO	0	64	<65	1
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	71	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	71	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	77	65-80	2
CDISCPILOT01	701	01-701-1023	XANOMELINE HIGH DOSE	81	77	65-80	2
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3
CDISCPILOT01	701	01-701-1047	PLACEBO	0	85	>80	3

Desperate Programmers

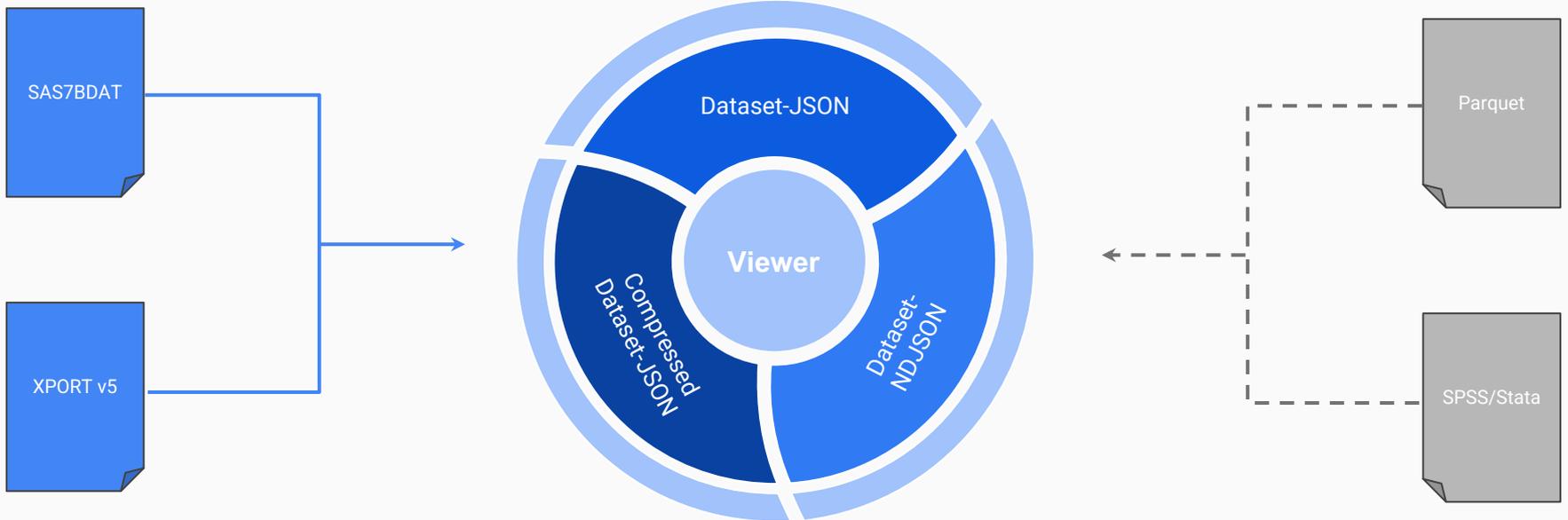


git



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 of rows and thousands of 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

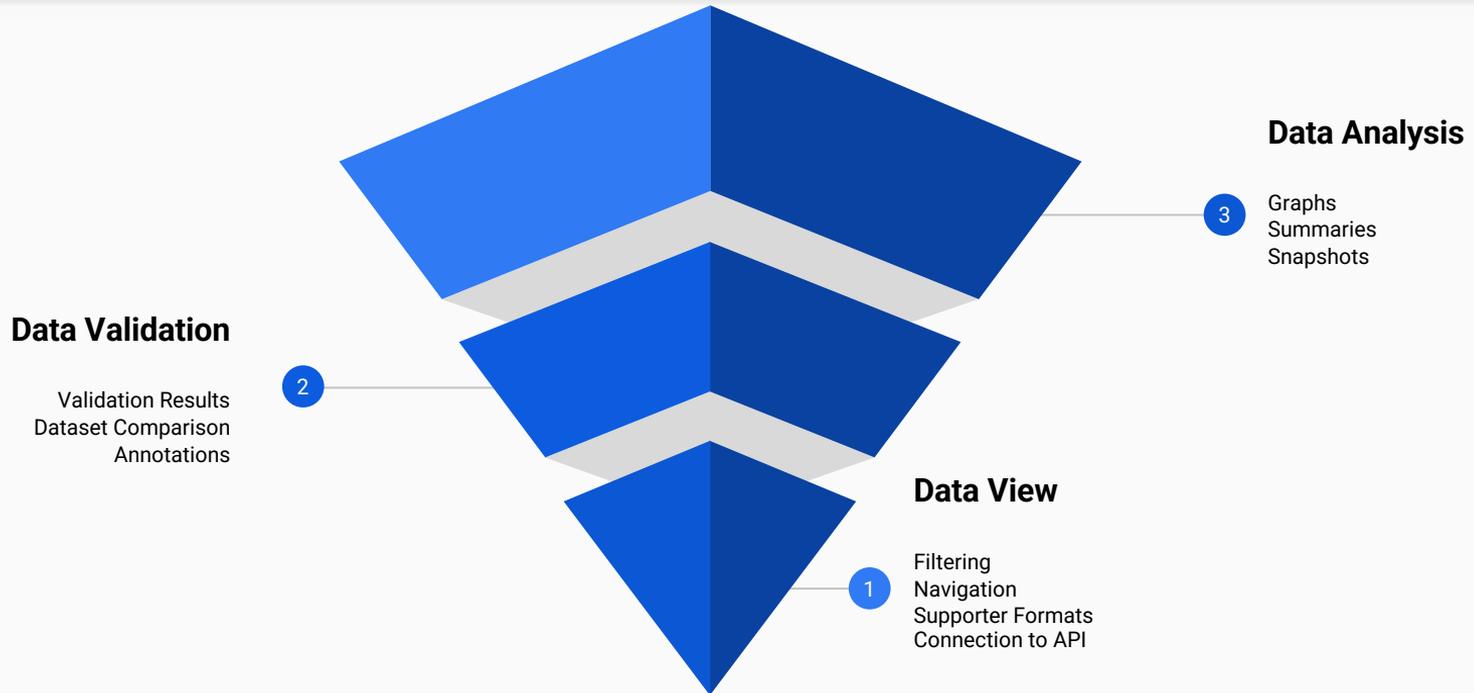
- Dataset-JSON API

- 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



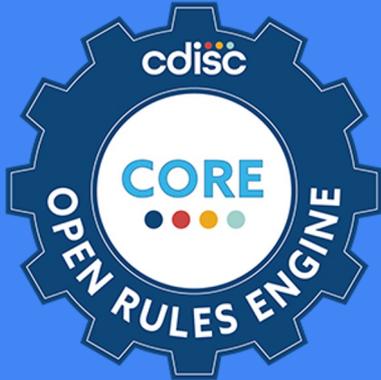
- 650+ employees
- Team is split over 60 countries
- Revenue over \$100 million



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>