

Single Day Event

**Next Phase of Data Analytics: How Graph
Databases Help with Metadata-Driven
Automation**

Tanisha & Vaibhav
GSK



Disclaimer

The views expressed herein are ours and don't represent GSK or any proprietary information of GSK.



Presenters & Presentation



With CDISC 360i, companies are collaborating on end-to-end automation, shifting to metadata-driven workflows that enhance efficiency and decision-making. GSK is pioneering a modular, automation-driven framework that accelerates clinical programming by breaking complex tasks into reusable modules - minimizing redundancy and enabling rapid adaptation.

Graph databases are key to this transformation, dynamically linking data collection, transformation, and CDISC-compliant outputs such as ADaM, Analysis Results Data, and TFLs. They strengthen metadata connectivity, enhance traceability, and support scalable automation.



Future expansions will extend this approach to SDTM, protocol development, and CSR writing - further improving accuracy, compliance, and efficiency. This presentation explores how metadata-driven automation, powered by graph databases, is shaping the next phase of data analytics and clinical reporting.



Outline

What is Metadata/ Graph databases

CDSIC 360i/Use case

Modularization, Revolutionization, Decoupling

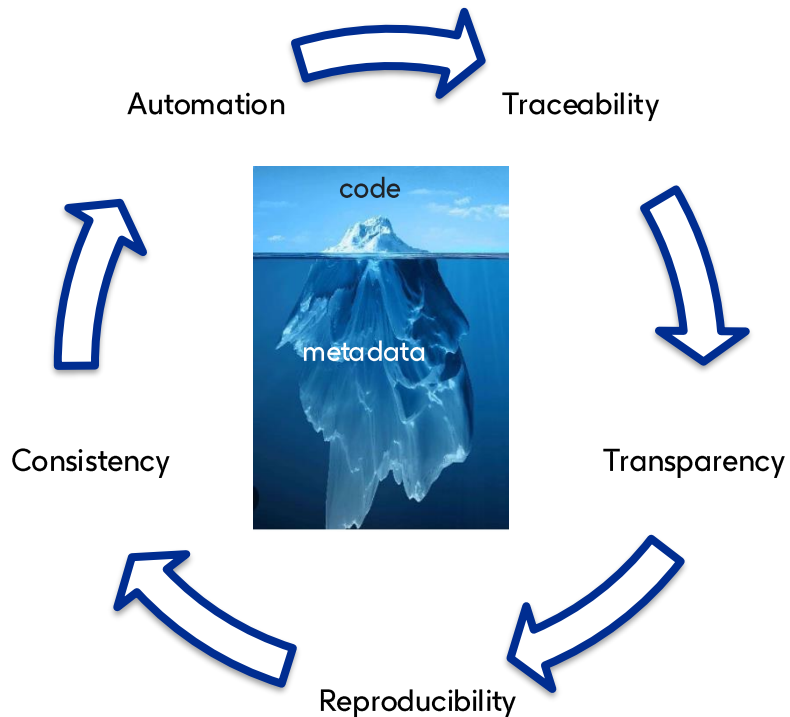
CDISC/company standards

Future Expectations

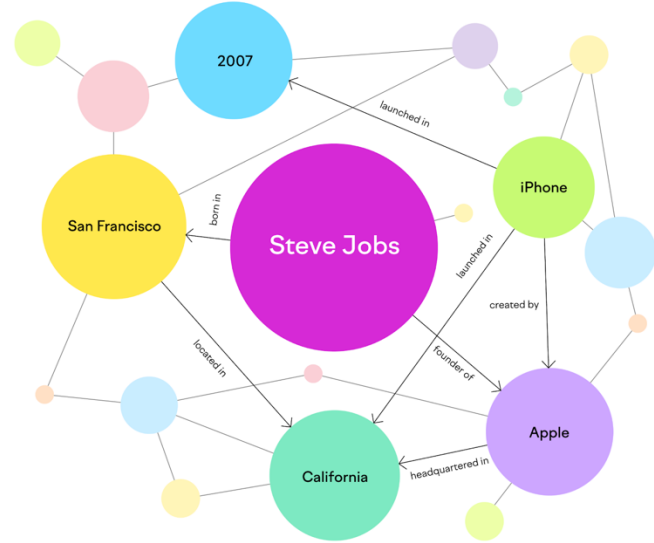
Conclusions



Metadata



The-old-way.xls

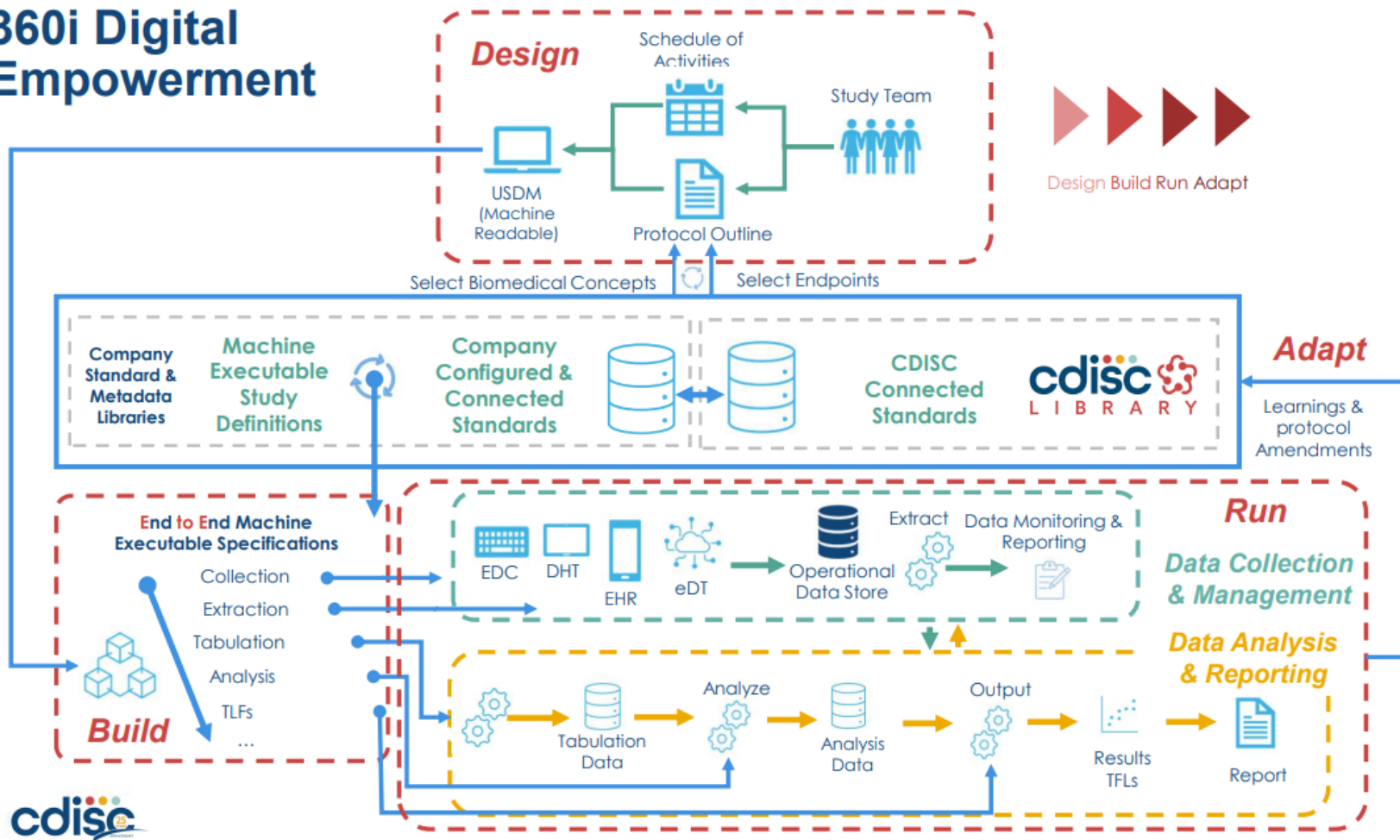


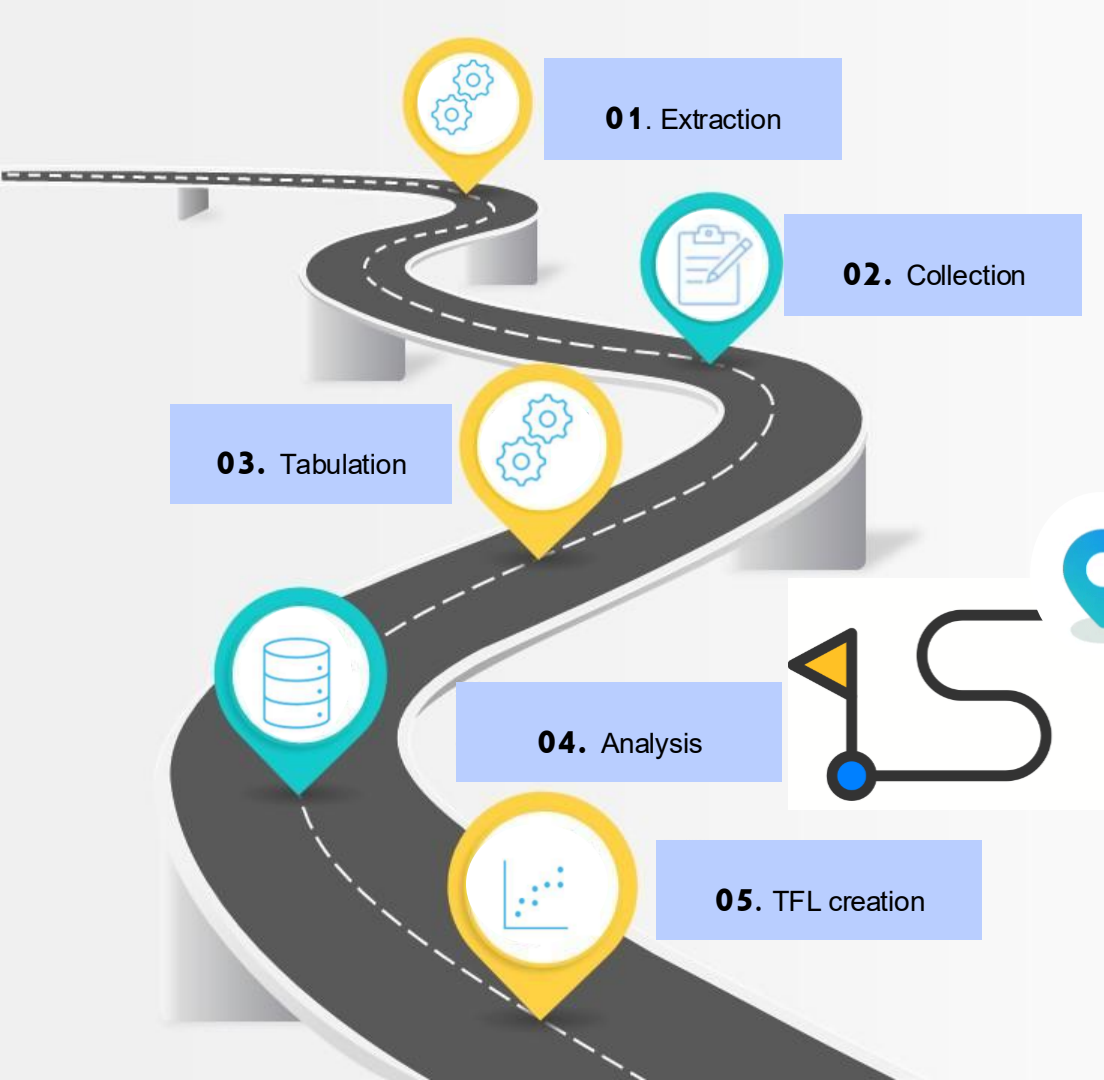


CDISC Vision



360i Digital Empowerment





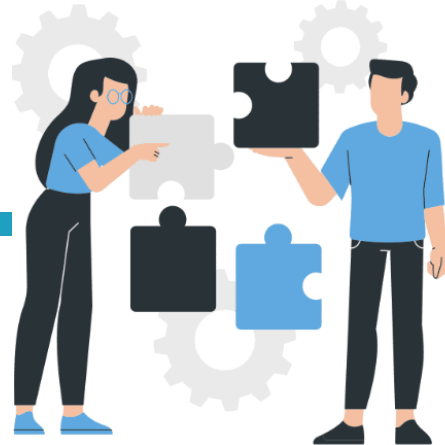
End-to-End standards

Conceptualize standards and solutions from end to end...trial design, collection, tabulation, analysis and reporting

We are here



Introduction: From Data to Decision



- Manual steps
- Resulting in inefficiencies and disconnects
- Leading to longer cycle times

Current Challenge

Our Approach

- A modular, automation-driven framework that integrates linked metadata, streamlining workflows from data collection to submission

- Faster, CDISC-compliant submissions, reduced cycle times

The Impact



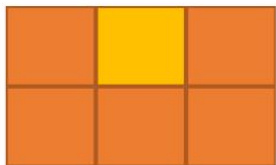


Monolith to Modules

TODAY

if an item has a single deviation from the standards, the entire deliverable is classified and treated as non-standard.

Non-Standard



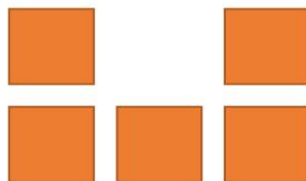
IN FUTURE STATE

with each component being a separate entity, a deviation from the standards is limited to what is actually deviating only, and the deliverable has a higher standards adherence.

Non-Standard



Standard



Modularization



Revolutionizing Clinical Programming: How do we define a module?

- Source Data - From?
- Code References - How?
- Output – To? [TREASURE]
- Description – Finding the treasure (Defining metadata to derive AAGE)

Collecting source data to derive AAGE

- LIBRARY: SDTM
- DATASET: DM
- VARIABLE: USUBJID, BRTHDTC
- FILTER: None



SOURCE DATA - FROM

References to the code to execute

PATH – HOW

- PROGRAMMING LANGUAGE: R
- PACKAGE: ADMIRAL
- FUNCTION: DERIVE AAGE
- PARAMETERS: ...

Adding AAGE to ADSL dataset

TO - TREASURE



- LIBRARY: ADAM
- DATASET: ADSL?
- VARIABLE: AAGE
- JOIN
- KEYS: STUDYID, USUBJID, SUBJID
- VALUE LEVEL?

TREASURE MAP – AAGE MODULE

- LIBRARY: ADAM, SDTM, TFL
- DATASET: DM
- VARIABLE: USUBJID, BRTHDTC
- FILTER



SOURCE DATA - FROM

PATH – HOW

- PROGRAMMING LANGUAGE: R
- PACKAGE: ADMIRAL
- FUNCTION: DERIVE AAGE
- PARAMETERS: ...

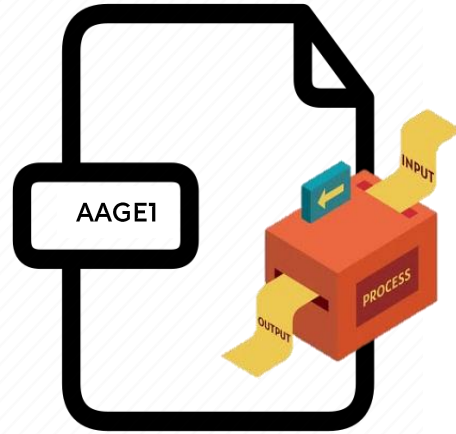
- LIBRARY: ADAM, SDTM, TFL
- DATASET: ADSL?
- VARIABLE: AAGE
- JOIN
- KEYS: STUDYID, USUBJID, SUBJID
- VALUE LEVEL?



TO – TREASURE

STANDARDIZED

HUMAN READABLE



INTEROPERABLE

STRUCTURED





CONNECTING THE DOTS



Built-in relationships



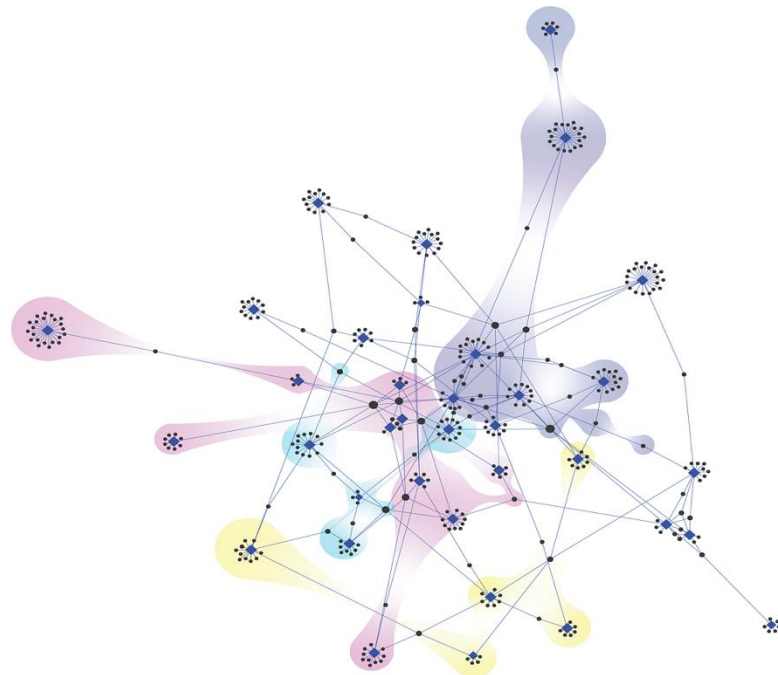
Flexibility: Schema-Free



Reuse

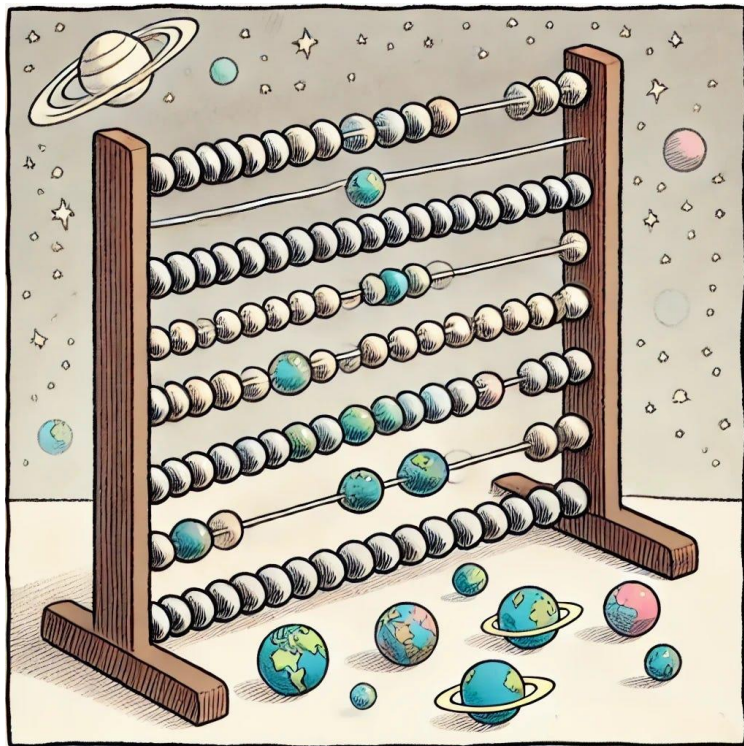


Execution Order





CDISC/Company Standards Multiverse



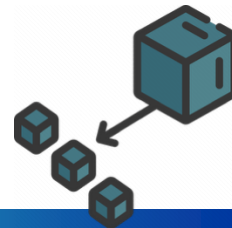
- e2e – Content v/s executable
- Mock shells
- CDISC compliance



Imagine how metadata can be defined in the form of modules in a TFL scenario!



Decoupling



A single analysis on one variable using one module

- 1 Frequency counts by treatment
- 2 Frequency counts by treatment for subjects with non-missing sex
- 3 Frequency counts by treatment and sex
- 4 Summary statistics on age by treatment

Reusability

How often do we report analysis 1. as big N?

How often do we use analysis 1. as denominator?

Should we generate analysis 1. every time we need it?

Protocol: ABC123456 Page 1 of 1
Analysis Set: Safety (Data as of: ddMMMyyyy)

Table X: Summary of Demographic and Baseline Characteristics

ADSLTRT01A		Treatment A	Treatment B	Total
		(N=xxx)	(N=xxx)	(N=xxx)
Sex	ADSLSEX	big N		
n		xxx	xxx	xxx
F		xx (xx%)	xx (xx%)	xx (xx%)
M		xx (xx%)	xx (xx%)	xx (xx%)
Age (YEARS) at First Dose [1]		ADSLAAGE		
n		xxx	xxx	xxx
Mean		xx.x	xx.x	xx.x
SD		x.xx	x.xx	x.xx
Median		xx.x	xx.x	xx.x
Min.		xx	xx	xx
Max.		xx	xx	xx

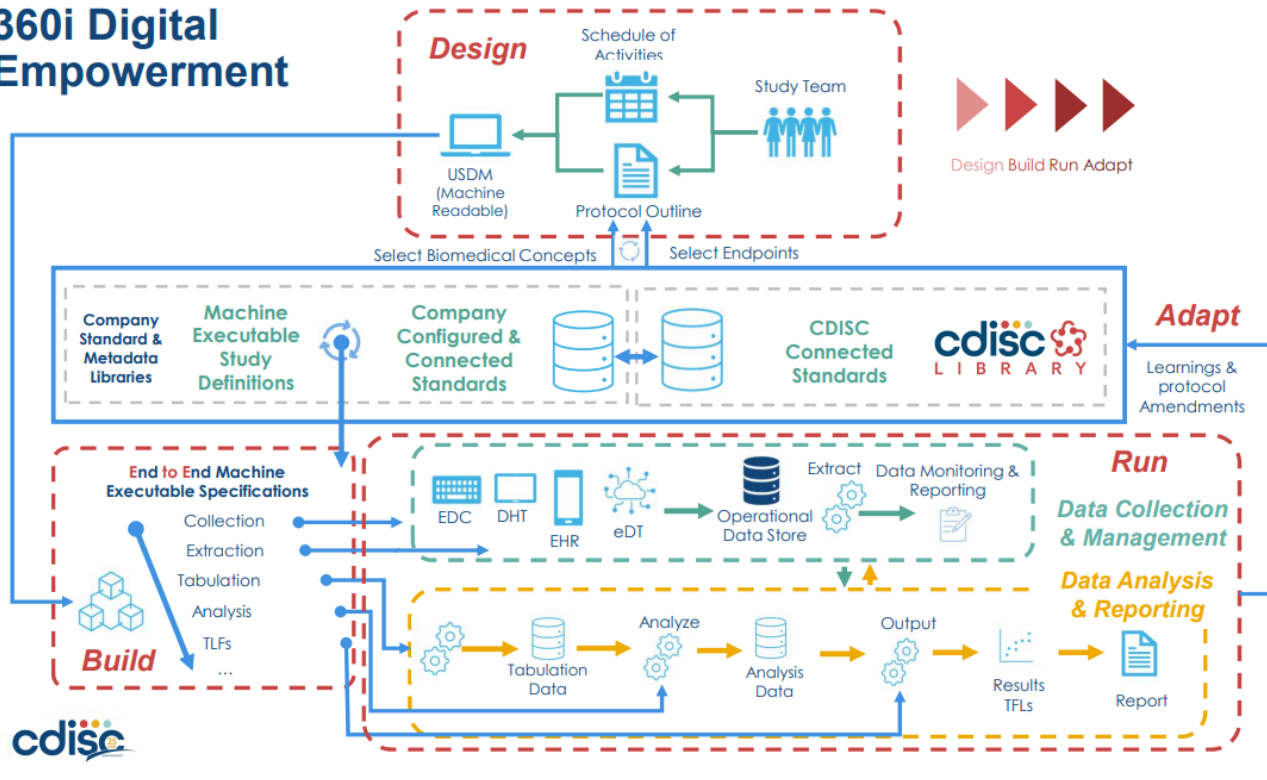
[1] Age is imputed when full date of birth is not provided.
/Directory/program.R ddMMMyyyy HH:mm





Unlocking the Full Power of Metadata

360i Digital Empowerment





- Accelerated research timelines.
- Automation of e2e standards implementation.
- Fosters collaboration and facilitates compliance with regulatory standards.
- Enhanced data integrity.
- Decoupling of analysis result production from static output.
- Improved reproducibility and adaptability in research endeavors.



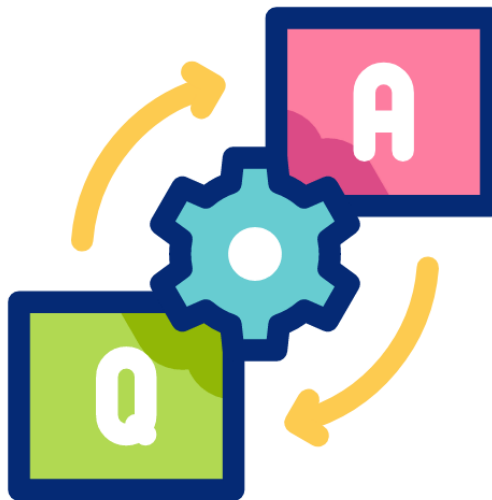
The Global Healthcare Data Science Community

Contact Channels

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

Social Media

@phusetwit
/phusebook
/phusetube
/company/phuse



* All the images used in this presentation have been sourced from Google & GSK brand hub.