



SpecToSAS

AI-Powered SAS Code Generation from Study Specs

Gargee Vaidya



Problem Statement

- **Manual Processes:** Programmers manually iterate through 100+ variables writing SAS code for each.
- **Time-Consuming:** Each variable has its own derivation rule making the process labor-intensive.
- **Productivity Impact:** Routine tasks consume valuable time that could be spent on quality control and analysis.

Dataset Name	Variable Name	Variable Label	Length	Control	CDISC Origin	AZ Core	Varial	AZ Derivation
ADAE	STUDYID	Study Identifier	21		Predecessor	Req	C	AE.STUDYID
ADAE	USUBJID	Unique Subject Identifier	30		Predecessor	Req	C	AE.USUBJID
ADAE	AESEQ	Sequence Number	8		Predecessor	Req	N	AESEQ
ADAE	AESPID	Sponsor-Defined Identifier	40		Predecessor	Perm	C	AE.AESPID
ADAE	AETERM	Reported Term for the Adverse	200		Predecessor	Req	C	AE.AETERM
ADAE	AELLT	Lowest Level Term	200	MedDRA	Predecessor	Cond	C	AE.AELLT
ADAE	AELLTCD	Lowest Level Term Code	8	MedDRA	Predecessor	Perm	N	AE.AELLTCD
ADAE	AEDECOD	Dictionary-Derived Term	200	MedDRA	Predecessor	Req	C	AE.AEDECOD
ADAE	AEPTCD	Preferred Term Code	8	MedDRA	Predecessor	Perm	N	AE.AEPTCD
ADAE	AEHLT	High Level Term	200	MedDRA	Predecessor	Cond	C	AE.AEHLT
ADAE	AEHLTCD	High Level Term Code	8	MedDRA	Predecessor	Perm	N	AE.AEHLTCD
ADAE	AEHLGT	High Level Group Term	200	MedDRA	Predecessor	Cond	C	AE.AEHLGT
ADAE	AEHLGTC	High Level Group Term Code	8	MedDRA	Predecessor	Perm	N	AE.AEHLGTC
ADAE	AEBO	Body System or Organ Class	200	MedDRA	Predecessor	Req	C	AE.AEBO
ADAE	AEBO	Body System or Organ Class Code	8	MedDRA	Predecessor	Perm	N	AE.AEBO
ADAE	AESOC	Primary System Organ Class	200	MedDRA	Predecessor	Cond	C	AE.AESOC
ADAE	AESOC	Primary System Organ Class Code	8	MedDRA	Predecessor	Perm	N	AE.AESOC
ADAE	AESEV	Severity/Intensity	40		Predecessor	Perm	C	AE.AESEV
ADAE	AESER	Serious Event	2 (NY)		Predecessor	Req	C	AE.AESER
ADAE	AEACN	Action Taken with Study Treatment	40		Predecessor	Perm	C	AE.AEACN
ADAE	AEACN	Other Action Taken	200		Predecessor	Perm	C	Describes other acti
ADAE	AEREL	Causality	200 (AEREL)		Predecessor	Perm	C	Records the investig
ADAE	AEOUT	Outcome of Adverse Event	200 (OUT)		Predecessor	Perm	C	Description of the ou
ADAE	AESDTH	Results in Death	1 (NY)		Predecessor	Perm	C	Did the serious event
ADAE	AECONTRT	Concomitant or Additional Trtmnt	1 (NY)		Predecessor	Perm	C	Was another treatme
ADAE	AETOXGR	Standard Toxicity Grade	40		Predecessor	Perm	C	AE.AETOXGR
ADAE	AESTDTC	Start Date/Time of Adverse Event	19 ISO 8601		Predecessor	Perm	C	Start date/time of the
ADAE	AEENDTC	End Date/Time of Adverse Event	19 ISO 8601		Predecessor	Cond	C	End date/time of the
ADAE	AESTDY	Study Day of Start of Adverse	8		Predecessor	Perm	N	Study day of start of
ADAE	AEENDY	Study Day of End of Adverse Event	8		Predecessor	Perm	N	Study day of end of
ADAE	AEDUR	Duration of Adverse Event	40 ISO 8601		Predecessor	Perm	C	Collected duration an
ADAE	AEACN1	Action Taken, Investigational	40		Predecessor	Perm	C	
ADAE	AEACNAD	Action Taken, Additional Drug	40		Predecessor	Perm	C	
ADAE	AEACNAD1	Action Taken, Additional Drug 1	40		Predecessor	Perm	C	
ADAE	AETRTEM	Treatment Emergent Flag	2 NY		Predecessor	Perm	C	SUPPAE.QVAL wher
ADAE	MEDDRAV	MedDRA version	40		Predecessor	Perm	C	SUPPAE.QVAL wher
ADAE	SAEDTC	Date/Time AE Met Criteria for	19 ISO8601		Predecessor	Perm	C	From SUPPAE
ADAE	ASEQ	Analysis Sequence Number	8		Derived	Perm	N	Sequence number g
ADAE	ARM	Description of Planned Arm	200		Predecessor	Perm	C	ADSL.ARM
ADAE	ARMCD	Planned Arm Code	20		Predecessor	Perm	C	ADSL.ARMCD



Introduction



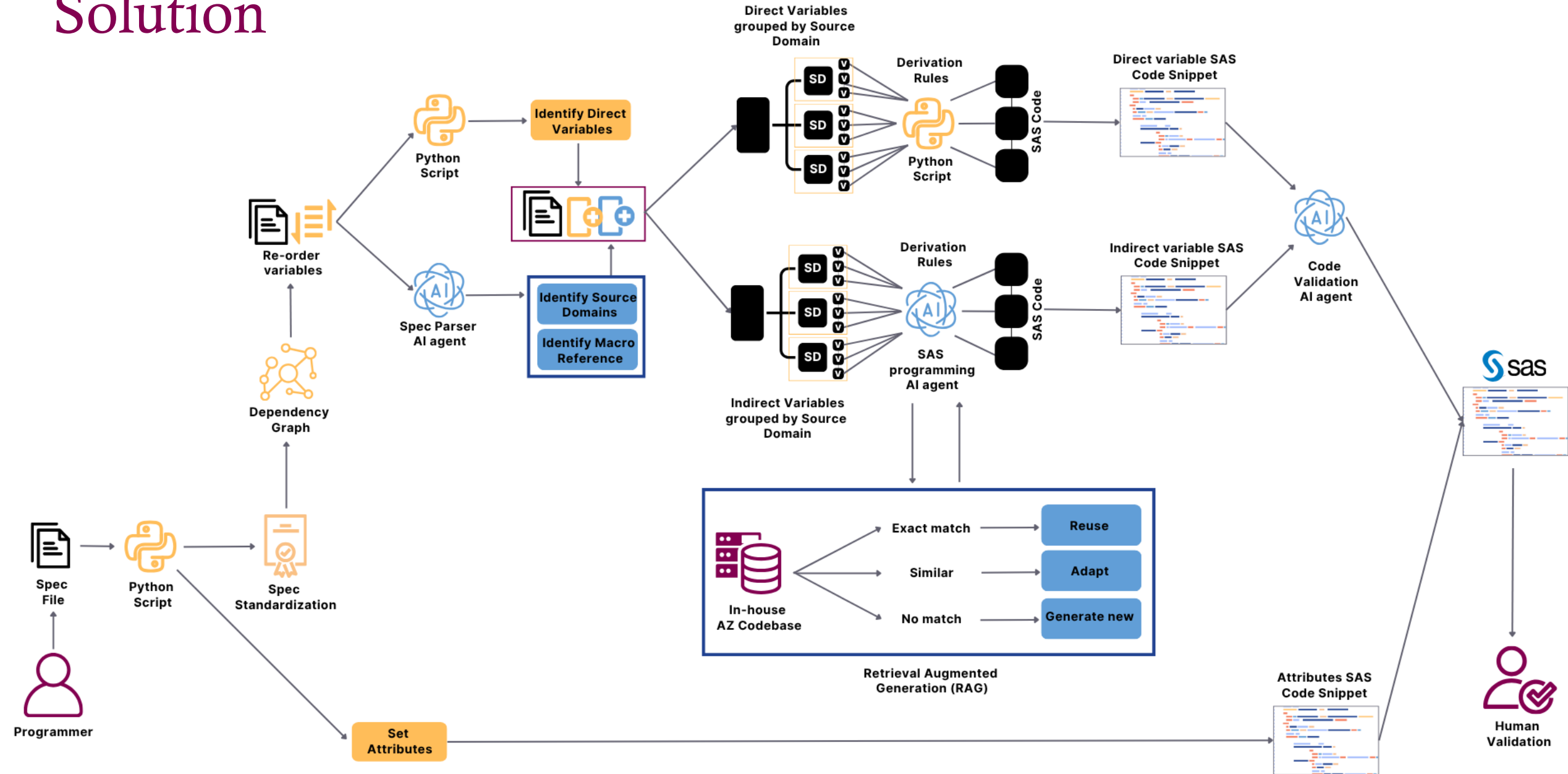
SpecToSas automates the process of generating SAS code for programming datasets using Generative AI and Python.



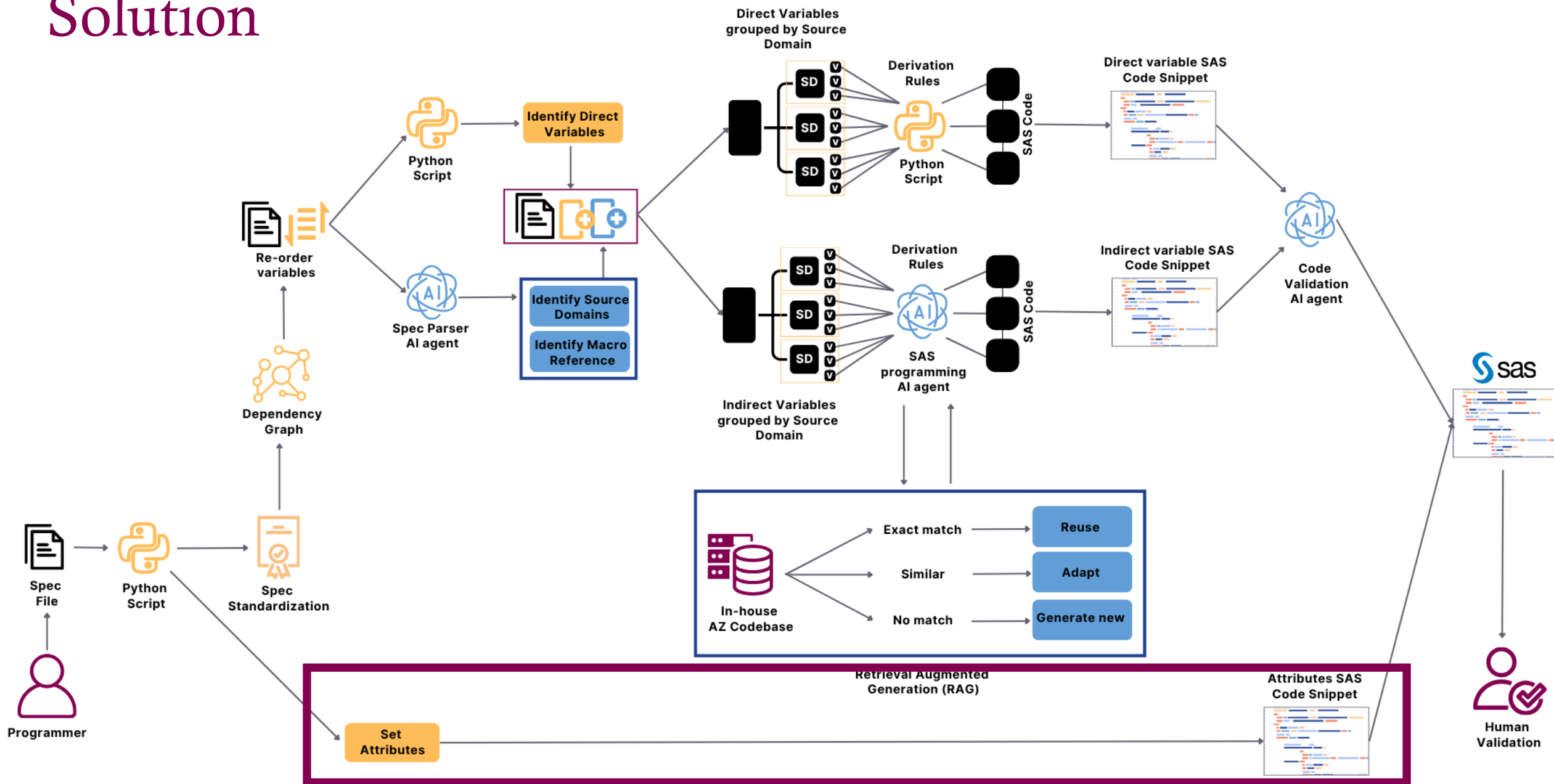
The tool helps programmers save time, reduce chance of errors, and allow them to focus on higher-value activities.



Solution



Solution



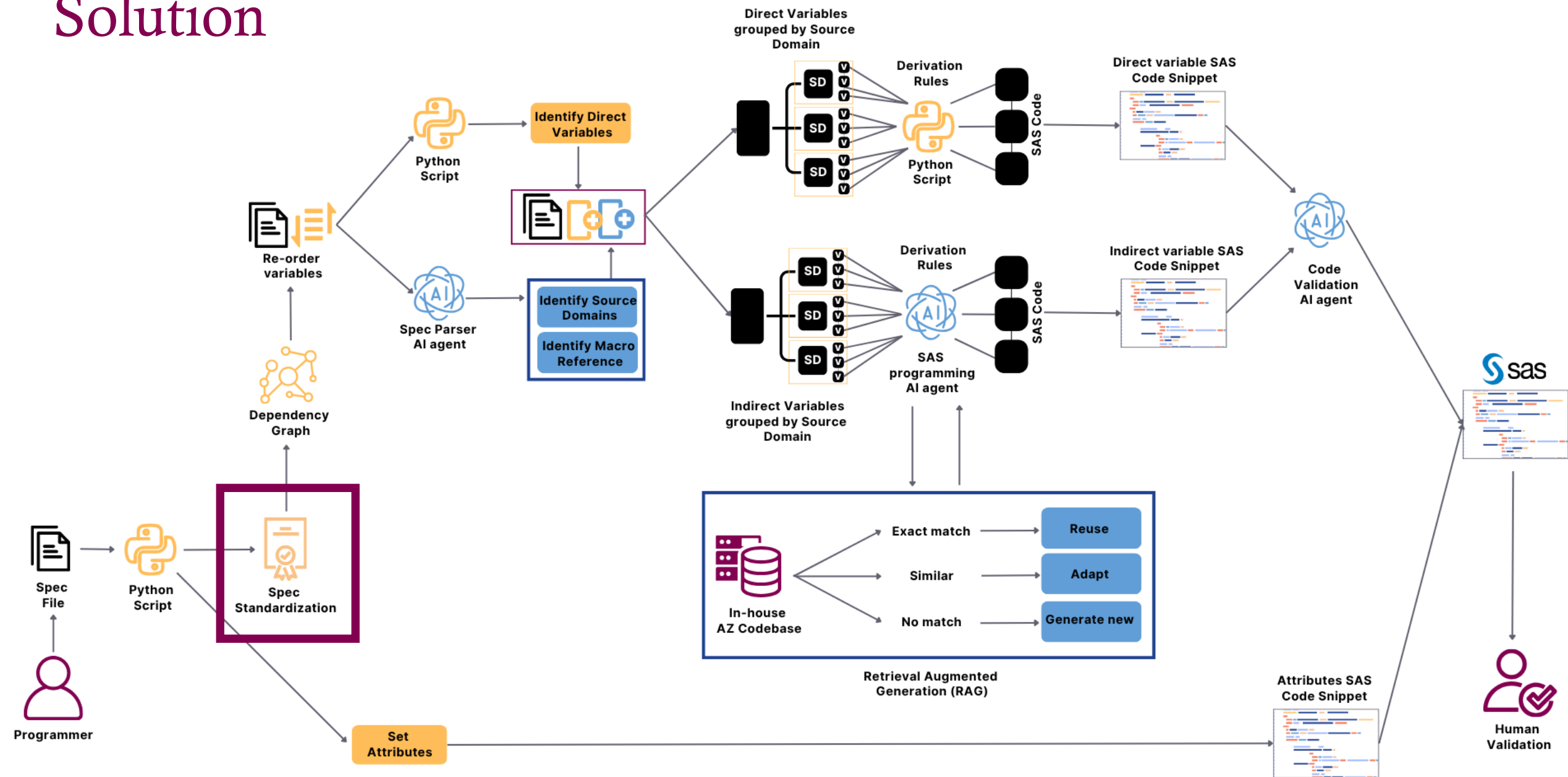
Set Attributes

Domain	Variable	Label	Type	Length	Origin
ADAE	STUDYID	Study Identifier	Char	200	ADSL.STUDYID
ADAE	USUBJID	Unique Subject Identifier	Char	200	ADSL.USUBJID
ADAE	SUBJID	Subject Identifier for the Study	Char	200	ADSL.SUBJID
ADAE	AESEQ	Sequence Number	Num	8	Sort by SUBJID AEDECOD ASTDT AETERM and assign sequential integer within each subject.
ADAE	ASTDT	Analysis Start Date	Num	8	1) If full date, convert date portion of AE.aestdat to numeric SAS date.
ADAE	AETERM	Reported Term for the Adverse Event	Char	200	Set to AE.aeterm value. Convert to upper case.
ADAE	AEDECOD	Dictionary-Derived Term	Char	200	Set to AE.pt_name value
ADAE	AESDTH	Results in Death	Char	200	SERAE.AESDTH. Merge SERAE to AE by subject and aeno. '1' = 'Y' and '0' = 'N'

```
proc sql;
  create table OUTPUT_FILENAME as
  select
    STUDYID length = 200 label = 'Study Identifier',
    USUBJID length = 200 label = 'Unique Subject Identifier',
    SUBJID length = 200 label = 'Subject Identifier for the Study',
    AESEQ label = 'Sequence Number',
    AETERM length = 200 label = 'Reported Term for the Adverse Event',
    AEDECOD length = 200 label = 'Dictionary-Derived Term',
    AESDTH length = 200 label = 'Results in Death',
    ASTDT label = 'Analysis Start Date',
  from INPUT_FILENAME;
quit;
```



Solution



Spec Standardization

Domain	Variable	Label	Type	Length	Origin
ADLB	CRIT1	Analysis Criterion 1	Char	200	See Value Level Metadata
ADLB	CRIT1FL	Criterion 1 Evaluation Result Flag	Char	2	See Value Level Metadata

Variable Metadata

Value Level Metadata

Domain	Parameter Identifier	Variable	Origin
ADLB	L55101U1	CRIT1	Set to 'ALT >= 3 x ULN'
ADLB	L55102U1	CRIT1	Set to 'AST >= 3 x ULN'
ADLB	L55110U1	CRIT1	Set to 'Bilirubin >= 2 x ULN'
ADLB	L55101U1	CRIT1FL	Set to 'Y' where ADLB.RANRHI >= 3 or (ADLB.R2A1HI >= 3 and ADLB.R2A1FL >= 3)
ADLB	L55102U1	CRIT1FL	Set to 'Y' where ADLB.RANRHI >= 3 or (ADLB.R2A1HI >= 3 and ADLB.R2A1FL >= 3)
ADLB	L55110U1	CRIT1FL	Set to 'Y' where ADLB.RANRHI >= 2 or (ADLB.R2A1HI >= 2 and ADLB.R2A1FL >= 2)

Variable Metadata

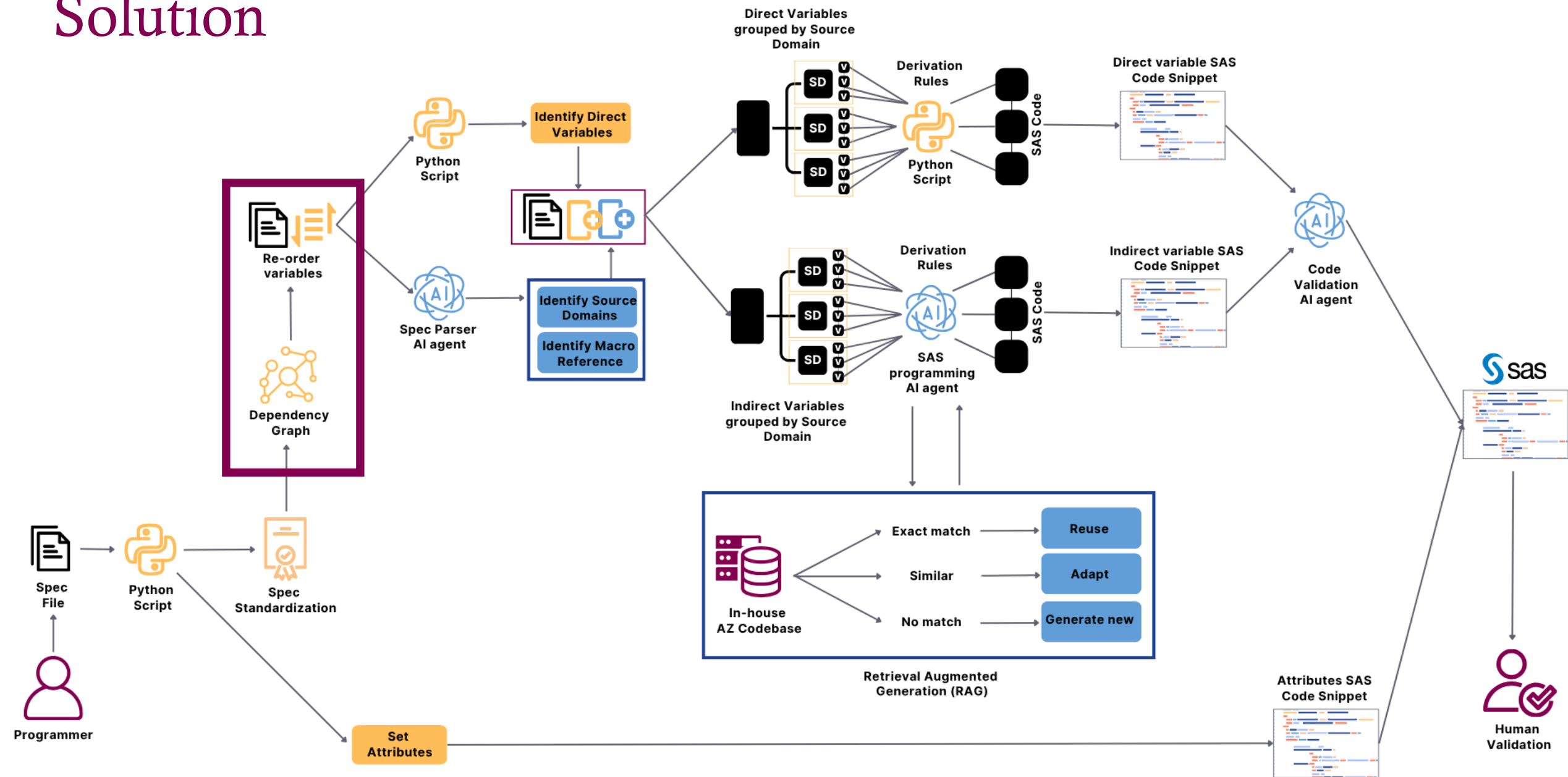
Value Level Metadata



Domain	Variable	Parameter Identifier	Label	Type	Length	Origin
ADLB	CRIT1	L55101U1	Analysis Criterion 1	Char	200	Set to 'ALT >= 3 x ULN'
ADLB	CRIT1	L55102U1	Analysis Criterion 1	Char	200	Set to 'AST >= 3 x ULN'
ADLB	CRIT1	L55110U1	Analysis Criterion 1	Char	200	Set to 'Bilirubin >= 2 x ULN'
ADLB	CRIT1FL	L55101U1	Criterion 1 Evaluation Result Flag	Char	2	Set to 'Y' where ADLB.RANRHI >= 3 or (ADLB.R2A1HI >= 3 and ADLB.R2A1FL >= 3)
ADLB	CRIT1FL	L55102U1	Criterion 1 Evaluation Result Flag	Char	2	Set to 'Y' where ADLB.RANRHI >= 3 or (ADLB.R2A1HI >= 3 and ADLB.R2A1FL >= 3)
ADLB	CRIT1FL	L55110U1	Criterion 1 Evaluation Result Flag	Char	2	Set to 'Y' where ADLB.RANRHI >= 2 or (ADLB.R2A1HI >= 2 and ADLB.R2A1FL >= 2)



Solution



10



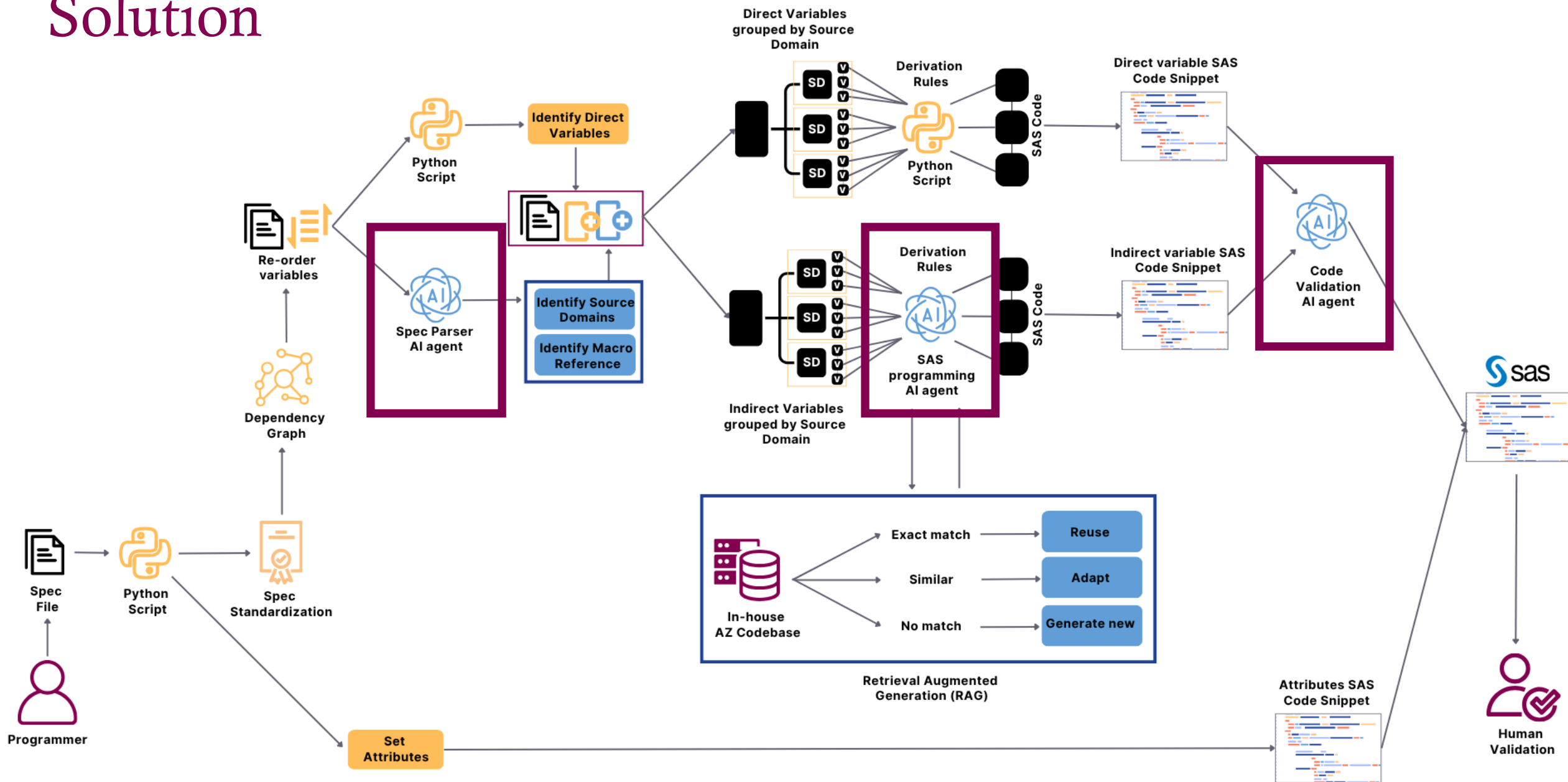
Reorder Variables

Domain	Variable	Label	Type	Length	Origin
ADAE	STUDYID	Study Identifier	Char	200	ADSL.STUDYID
ADAE	USUBJID	Unique Subject Identifier	Char	200	ADSL.USUBJID
ADAE	SUBJID	Subject Identifier for the Study	Char	200	ADSL.SUBJID
ADAE	AESEQ	Sequence Number	Num	8	Sort by SUBJID AEDECOD ASTDT AETERM and assign sequential integer within each subject.
ADAE	ASTDT	Analysis Start Date	Num	8	1) If full date, convert date portion of AE.aestdat to numeric SAS date.
ADAE	AETERM	Reported Term for the Adverse Event	Char	200	Set to AE.aeterm value. Convert to upper case.
ADAE	AEDECOD	Dictionary-Derived Term	Char	200	Set to AE.pt_name value
ADAE	AESDTH	Results in Death	Char	200	SERAE.AESDTH. Merge SERAE to AE by subject and aeno. '1' = 'Y' and '0' = 'N'

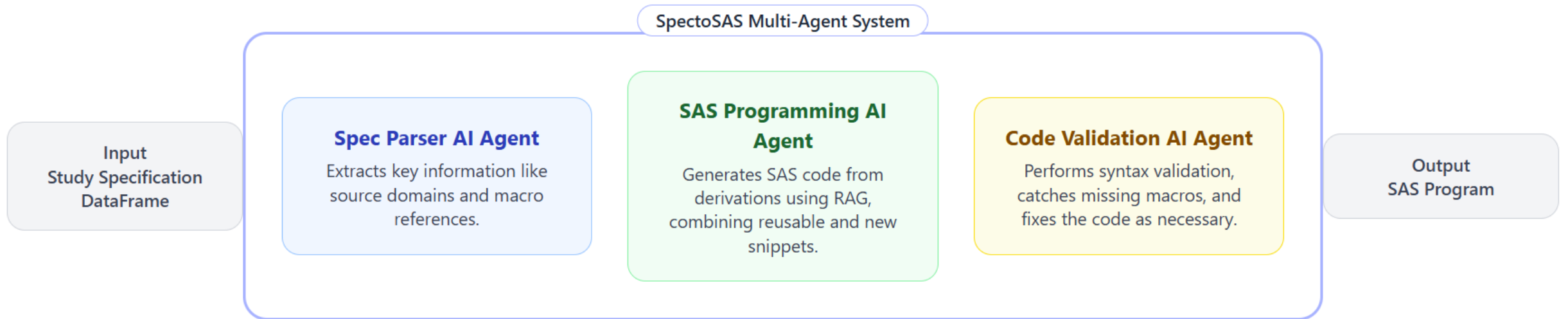
Domain	Variable	Label	Type	Length	Origin
ADAE	STUDYID	Study Identifier	Char	200	ADSL.STUDYID
ADAE	USUBJID	Unique Subject Identifier	Char	200	ADSL.USUBJID
ADAE	SUBJID	Subject Identifier for the Study	Char	200	ADSL.SUBJID
ADAE	ASTDT	Analysis Start Date	Num	8	1) If full date, convert date portion of AE.aestdat to numeric SAS date.
ADAE	AETERM	Reported Term for the Adverse Event	Char	200	Set to AE.aeterm value. Convert to upper case.
ADAE	AEDECOD	Dictionary-Derived Term	Char	200	Set to AE.pt_name value
ADAE	AESDTH	Results in Death	Char	200	SERAE.AESDTH. Merge SERAE to AE by subject and aeno. '1' = 'Y' and '0' = 'N'
ADAE	AESEQ	Sequence Number	Num	8	Sort by SUBJID AEDECOD ASTDT AETERM and assign sequential integer within each subject.



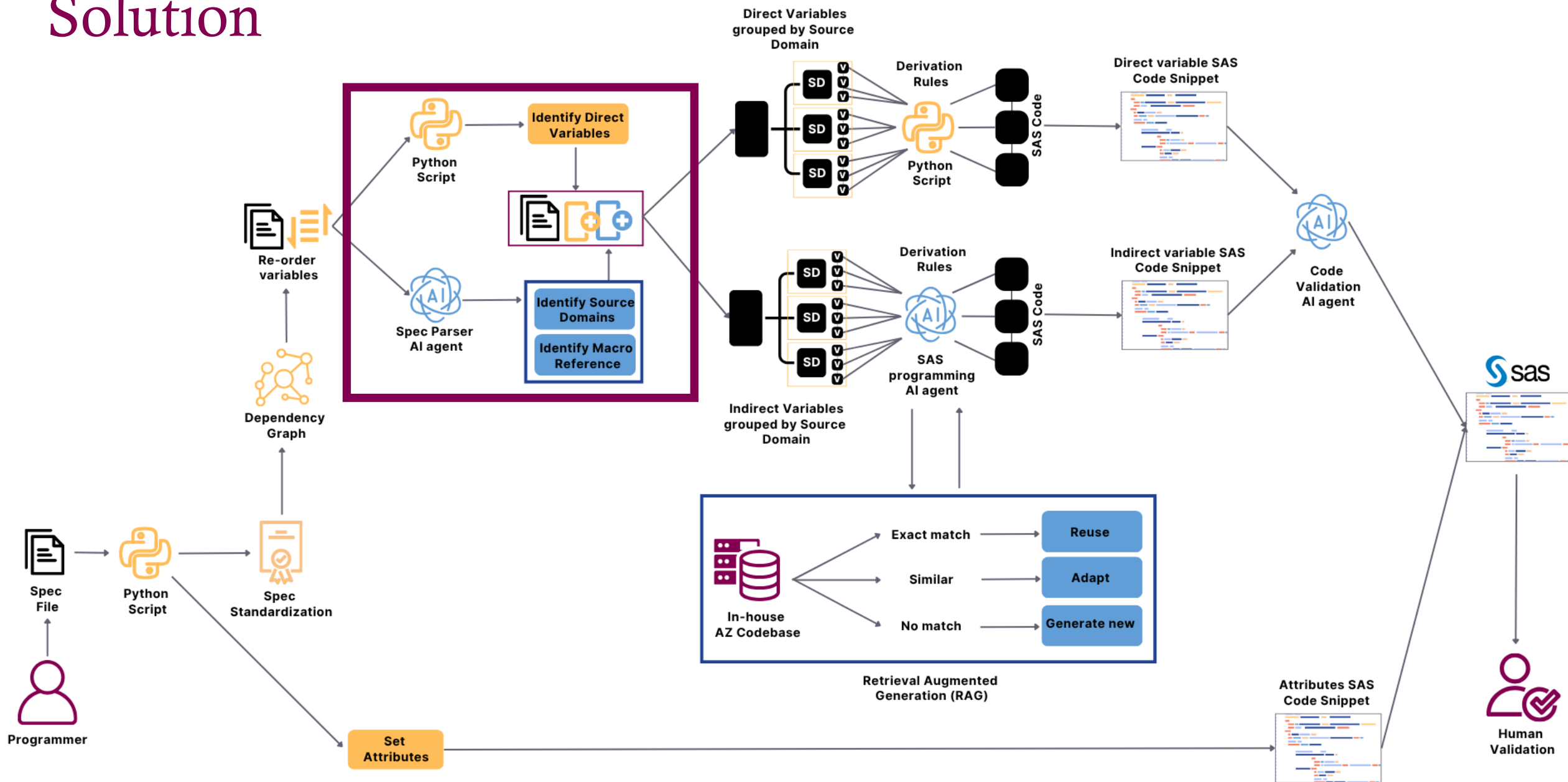
Solution



Multi-Agentic System Architecture



Solution



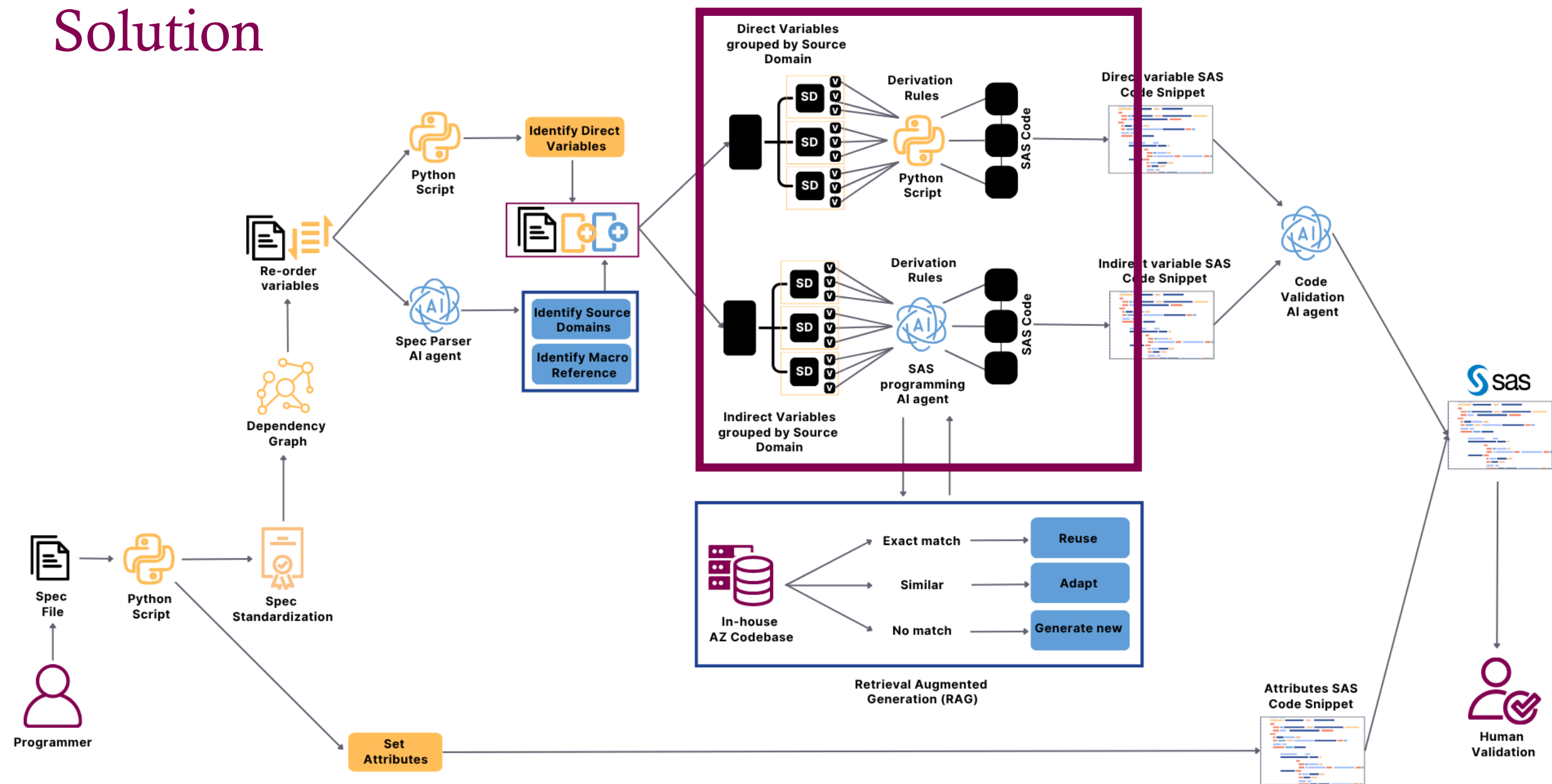
Direct Variables and Source Domain



Domain	Variable	Label	Type	Length	Origin	Direct Variable	Source Domain	Macro Reference
ADAE	STUDYID	Study Identifier	Char	200	ADSL.STUDYID	Y	ADSL	
ADAE	USUBJID	Unique Subject Identifier	Char	200	ADSL.USUBJID	Y	ADSL	
ADAE	SUBJID	Subject Identifier for the Study	Char	200	ADSL.SUBJID	Y	ADSL	
ADAE	ASTDT	Analysis Start Date	Num	8	1) If full date, convert date portion of AE.aestdat to numeric SAS date. 2) if partial date then do ; a) Missing day- Impute the 1st of the month unless month is the same as month of the first dose of study drug then impute first dose date b) Missing day and month -Impute 1st January unless year is the same as first dose date then impute first dose date c) Completely missing-Impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date.	N	AE	
ADAE	AETERM	Reported Term for the Adverse Event	Char	200	Set to AE.aeterm value. Convert to upper case.	N	AE	
ADAE	AEDECOD	Dictionary Derived Term	Char	200	Set to AE.pt_name value	N	AE	
ADAE	AESDTH	Results in Death	Char	200	SERAE.AESDTH. Merge SERAE to AE by subject and aeno.'1' = 'Y' and '0' = 'N'	N	AE, SERAE	
ADAE	MEDDRAV	MedDRA Version	Char	200	SUPPAE.QVAL where SUPPAE.QNAM = "MEDDRAV"	N	SUPPAE	%m_a_addsupp (dsin=)
ADAE	AESEQ	Sequence Number	Num	8	Sort by SUBJID AEDECOD ASTDT AETERM AESPID and assign sequential integer within each subject.	N	ADAE	



Solution



Code Generation

Variable	Label	Origin	Direct Variable	Source Domain
STUDYID	Study Identifier	ADSL.STUDYID	Y	ADSL
USUBJID	Unique Subject Identifier	ADSL.USUBJID	Y	ADSL
SUBJID	Subject Identifier for the Study	ADSL.SUBJID	Y	ADSL



```
data ADSL_;
  set ADSL;
  keep STUDYID USUBJID SUBJID;
run;
```

Variable	Label	Origin	Direct Variable	Source Domain
ASTDT	Analysis Start Date	1) If full date, convert date portion of AE.aestdat to numeric SAS date. 2) if partial date then do ; a) Missing day- Impute the 1st of the month unless month is the same as month of the first dose of study drug then impute first dose date b) Missing day and month -Impute 1st January unless year is the same as first dose date then impute first dose date c) Completely missing-Impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date	N	AE
AETERM	Reported Term for the Adverse Event	Set to AE.aeterm value. Convert to upper case.	N	AE
AEDECOD	Dictionary-Derived Term	Set to AE.pt_name value	N	AE
AESDTH	Results in Death	SERAE.AESDTH. Merge SERAE to AE by subject and aeno. '1' = 'Y' and '0' = 'N'	N	AE, SERAE
AESEQ	Sequence Number	Sort by SUBJID AEDECOD ASTDT AETERM and assign sequential integer within each subject.	N	ADAE



```
data AE_;
  set AE;
  AETERM = upcase(aeterm);
  AEDECOD = pt_name;

  format ASTDT date9.;

  if not missing(aestdat) then do;
    if intnx('year', aestdat, 0, 's') = aestdat then do;
      ASTDT = trtsdt;
    end;
    else if missing(day(aestdat)) then do;
      if month(aestdat) = month(trtsdt) then ASTDT = trtsdt;
      else ASTDT = mdy(month(aestdat), 1, year(aestdat));
    end;
    else if missing(day(aestdat)) and missing(month(aestdat)) then do;
      if year(aestdat) = year(trtsdt) then ASTDT = trtsdt;
      else ASTDT = mdy(1, 1, year(aestdat));
    end;
    else ASTDT = aestdat;
  end;
  else do;
    if year(aestdat) < year(trtsdt) then ASTDT = mdy(1, 1, year(aestdat));
    else ASTDT = mdy(1, 1, year(aestdat));
  end;
run;
```



Code Generation

Variable	Label	Origin	Direct Variable	Source Domain
ASTDT	Analysis Start Date	1) If full date, convert date portion of AE.aestdat to numeric SAS date. 2) if partial date then do ; a) Missing day- Impute the 1st of the month unless month is the same as month of the first dose of study drug then impute first dose date b) Missing day and month -Impute 1st January unless year is the same as first dose date then impute first dose date c) Completely missing-Impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date	N	AE
AETERM	Reported Term for the Adverse Event	Set to AE.aeterm value. Convert to upper case.	N	AE
AEDECOD	Dictionary-Derived Term	Set to AE.pt_name value	N	AE
AESDTH	Results in Death	SERAE.AESDTH. Merge SERAE to AE by subject and aeno. '1' = 'Y' and '0' = 'N'	N	AE, SERAE
AESEQ	Sequence Number	Sort by SUBJID AEDECOD ASTDT AETERM and assign sequential integer within each subject.	N	ADAE

```

data AE_SERAE_;
    merge AE(in=a) SERAE(in=b);
    by subject aeno;
    if b;
    format AESDTH $1.;
    if AESDTH = '1' then AESDTH = 'Y';
    else if AESDTH = '0' then AESDTH = 'N';
run;
    
```

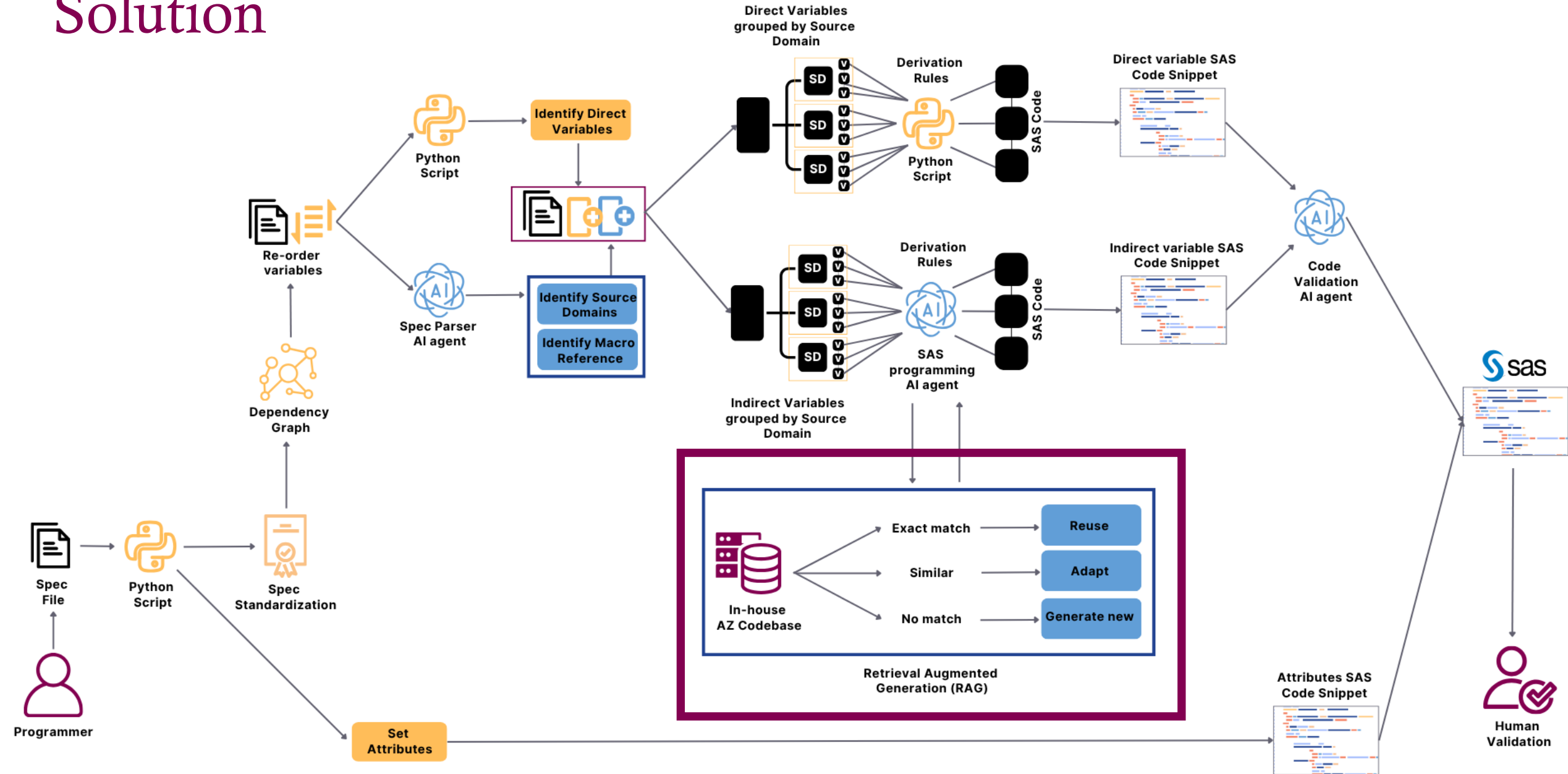


```

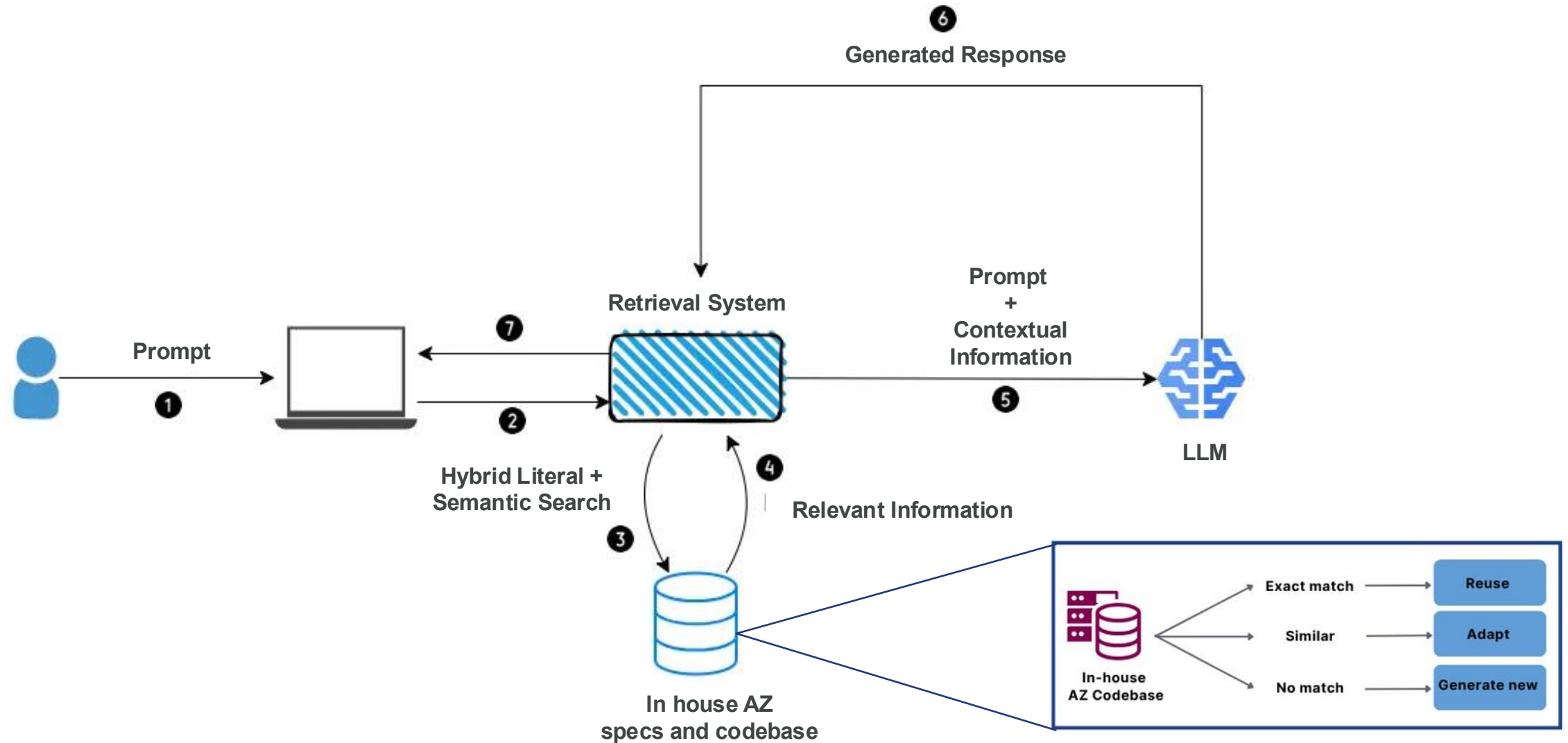
data ADAE;
    set ADAE_;
    by SUBJID AEDECOD ASTDT AETERM;
    if first.SUBJID then AESEQ = 1;
    else AESEQ + 1;
run;
    
```



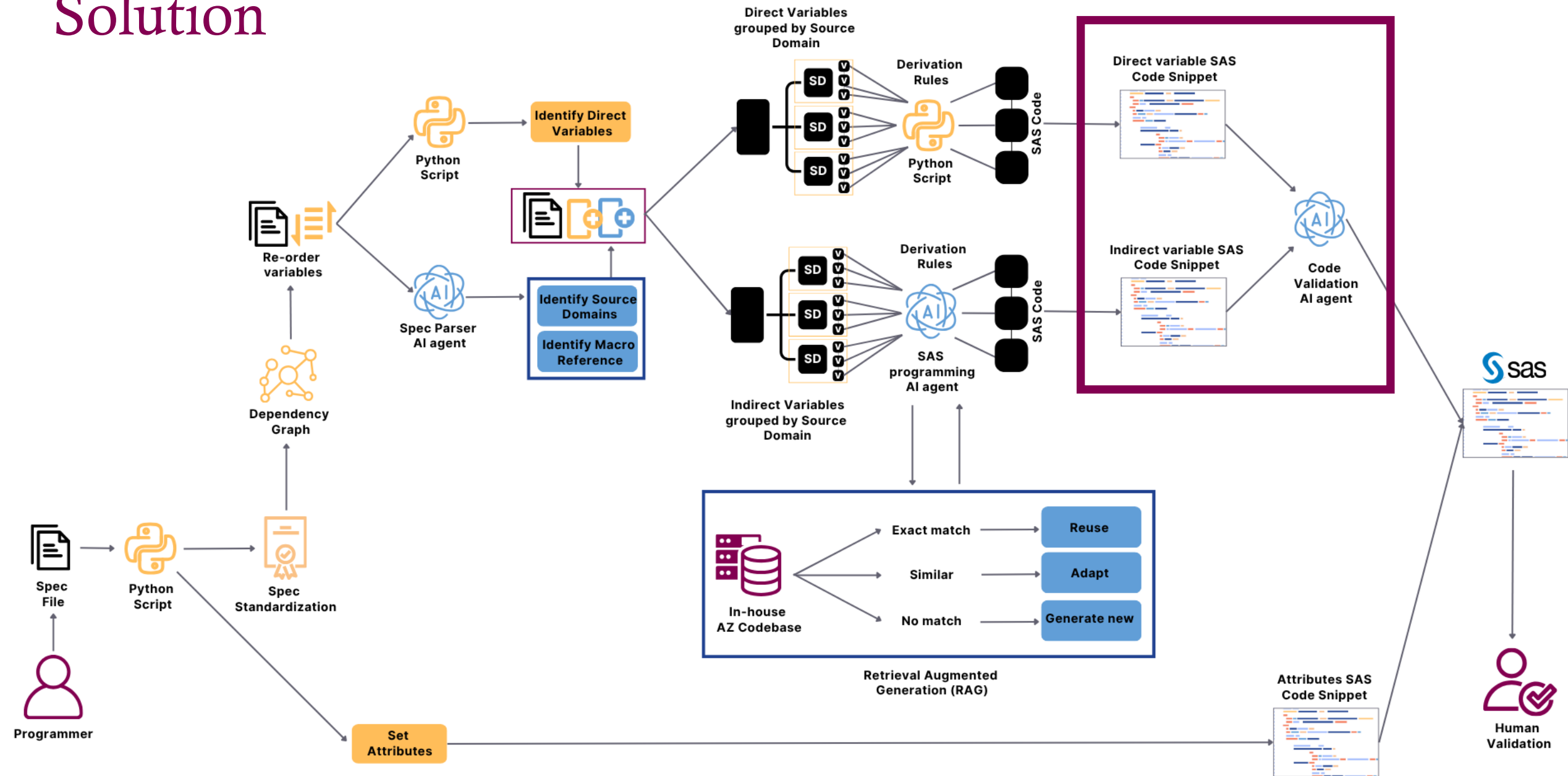
Solution



Retrieval Augmented Generation (RAG)



Solution



Code Generation

```
data ADSL_;  
  set ADSL;  
  keep STUDYID USUBJID SUBJID;  
run;
```

```
data AE_;  
  set AE;  
  AETERM = upcase(aeterm);  
  AEDECOD = pt_name;  
  
  format ASTDT date9.;  
  
  if not missing(aestdat) then do;  
    if intnx('year', aestdat, 0, 's') = aestdat then do;  
      ASTDT = trtsdt;  
    end;  
    else if missing(day(aestdat)) then do;  
      if month(aestdat) = month(trtsdt) then ASTDT = trtsdt;  
      else ASTDT = mdy(month(aestdat), 1, year(aestdat));  
    end;  
    else if missing(day(aestdat)) and missing(month(aestdat)) then do;  
      if year(aestdat) = year(trtsdt) then ASTDT = trtsdt;  
      else ASTDT = mdy(1, 1, year(aestdat));  
    end;  
    else ASTDT = aestdat;  
  end;  
end;  
if year(aestdat) < year(trtsdt) then ASTDT = mdy(1, 1, year(aestdat));  
else ASTDT = mdy(1, 1, year(aestdat));  
end;  
run;
```

```
data AE_SERAE_;  
  merge AE(in=a) SERAE(in=b);  
  by subject aeno;  
  if b;  
  format AESDTH $1.;  
  if AESDTH = '1' then AESDTH = 'Y';  
  else if AESDTH = '0' then AESDTH = 'N';  
run;
```

```
data ADAE;  
  set ADAE_;  
  by SUBJID AEDECOD ASTDT AETERM;  
  if first.SUBJID then AESEQ = 1;  
  else AESEQ + 1;  
run;
```

```
proc sql;  
  create table OUTPUT_FILENAME as  
  select  
    STUDYID length = 200 label = 'Study Identifier',  
    USUBJID length = 200 label = 'Unique Subject Identifier',  
    SUBJID length = 200 label = 'Subject Identifier for the Study',  
    AESEQ label = 'Sequence Number',  
    AETERM length = 200 label = 'Reported Term for the Adverse Event',  
    AEDECOD length = 200 label = 'Dictionary-Derived Term',  
    AESDTH length = 200 label = 'Results in Death',  
    ASTDT label = 'Analysis Start Date',  
  from INPUT_FILENAME;  
quit;
```

Direct variables
SAS Code



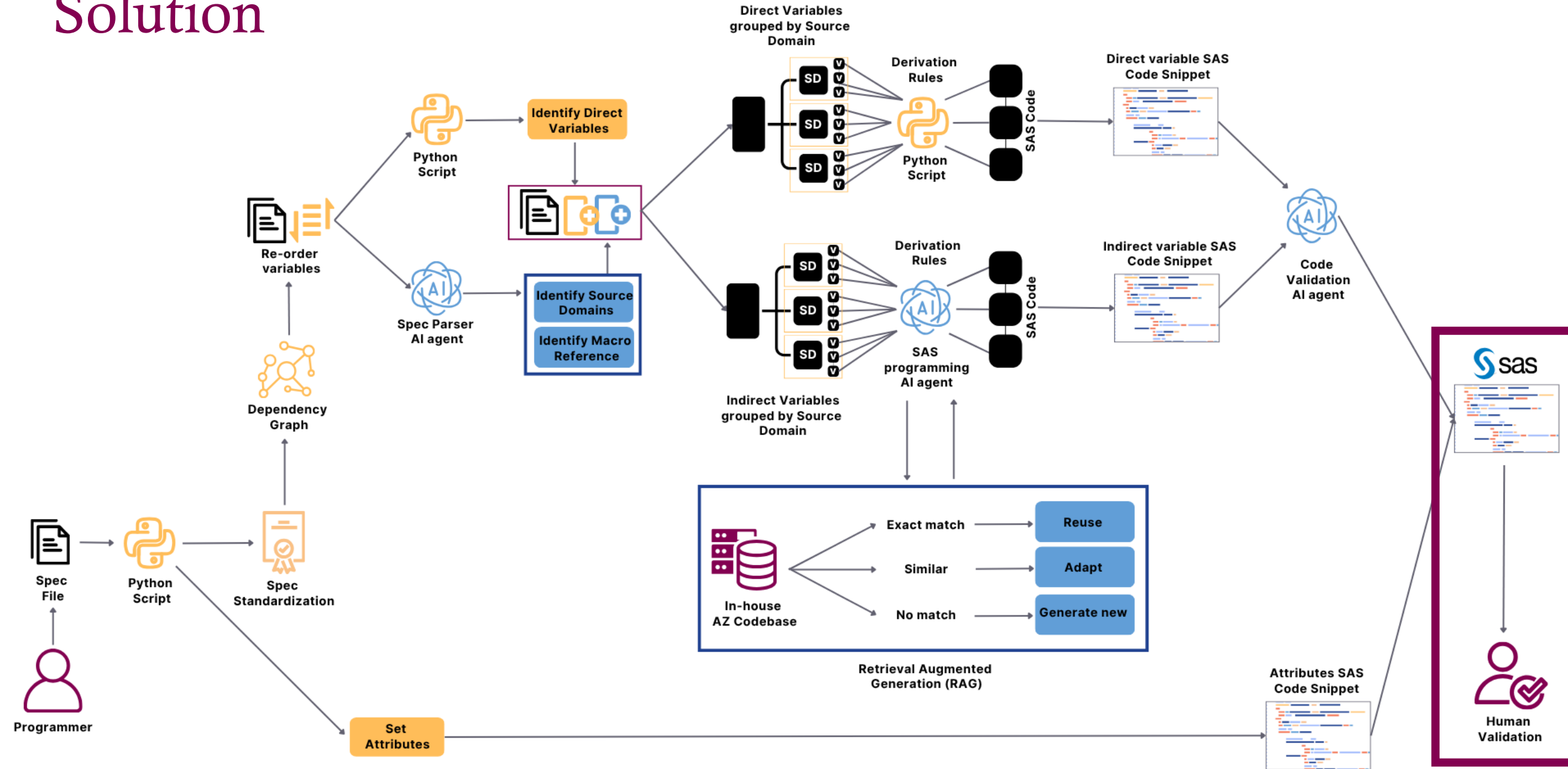
Indirect variables
SAS Code



Attributes
SAS Code



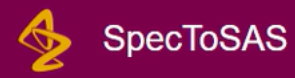
Solution



Demo



About



AZ Macros

Click to upload a document

```
1 // Your generated SAS code will appear here...
```

Copy Complete Code

ADAE	Y	SUBJID	Subject Identifier for the Study	Char	200		ADSL.SUB JID		Req	3.0	
ADAE	Y	SITEID	Study Site Identifier	Char	200		ADSL.SIT EID		Req	4.0	
ADAE	Y	COUNTR Y	Country	Char	200		ADSL.CO UNTRY		Req	5.0	

SAS code generation completed

Derived 47 of 47 variables...
Took 5 min 2 sec to generate SAS code.

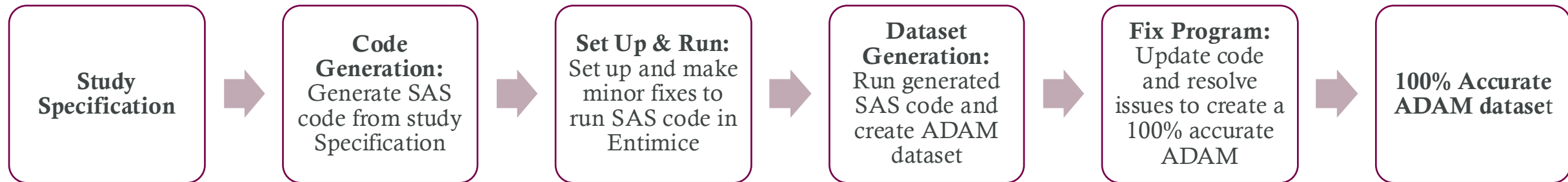
```
1  ***sas
2  /*****
3  * Program name      : program_name.sas
4  *
5  * Program path     : program path
6  *
7  * Type              : Output
8  *
9  * Purpose           :
10 *
11 * Author            :
12 *
13 * Date created      :
14 *
15 * Input datasets    :
16 *
17 * Macros used       :
18 *
19 * Output files      :
20 *
21 * Revision history  :
22 *
23 * Date      Author      Ref      Revision
24 *****/
25
26 ** Amend localsetup parameter so it points for the localsetup.sas file in your local area if needed **;
```

Copy Complete Code

Testing



Testing Process & Metrics




Code Generation:
5 to 15 mins


Set Up Time:
15 to 40 mins

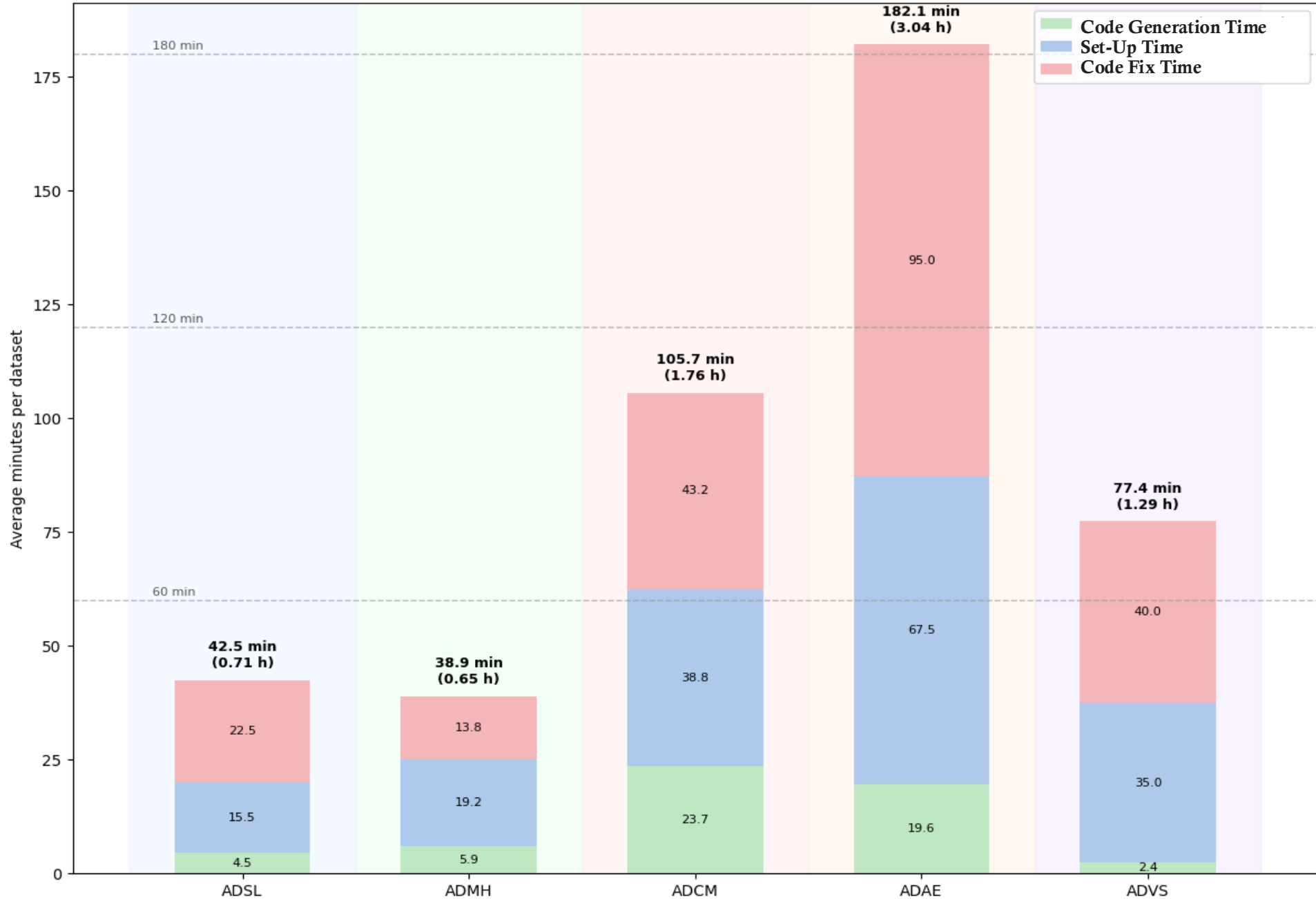

% Variable Derived:
75% to 95%


Code Fix Time
(100% accuracy):
25 to 55 mins

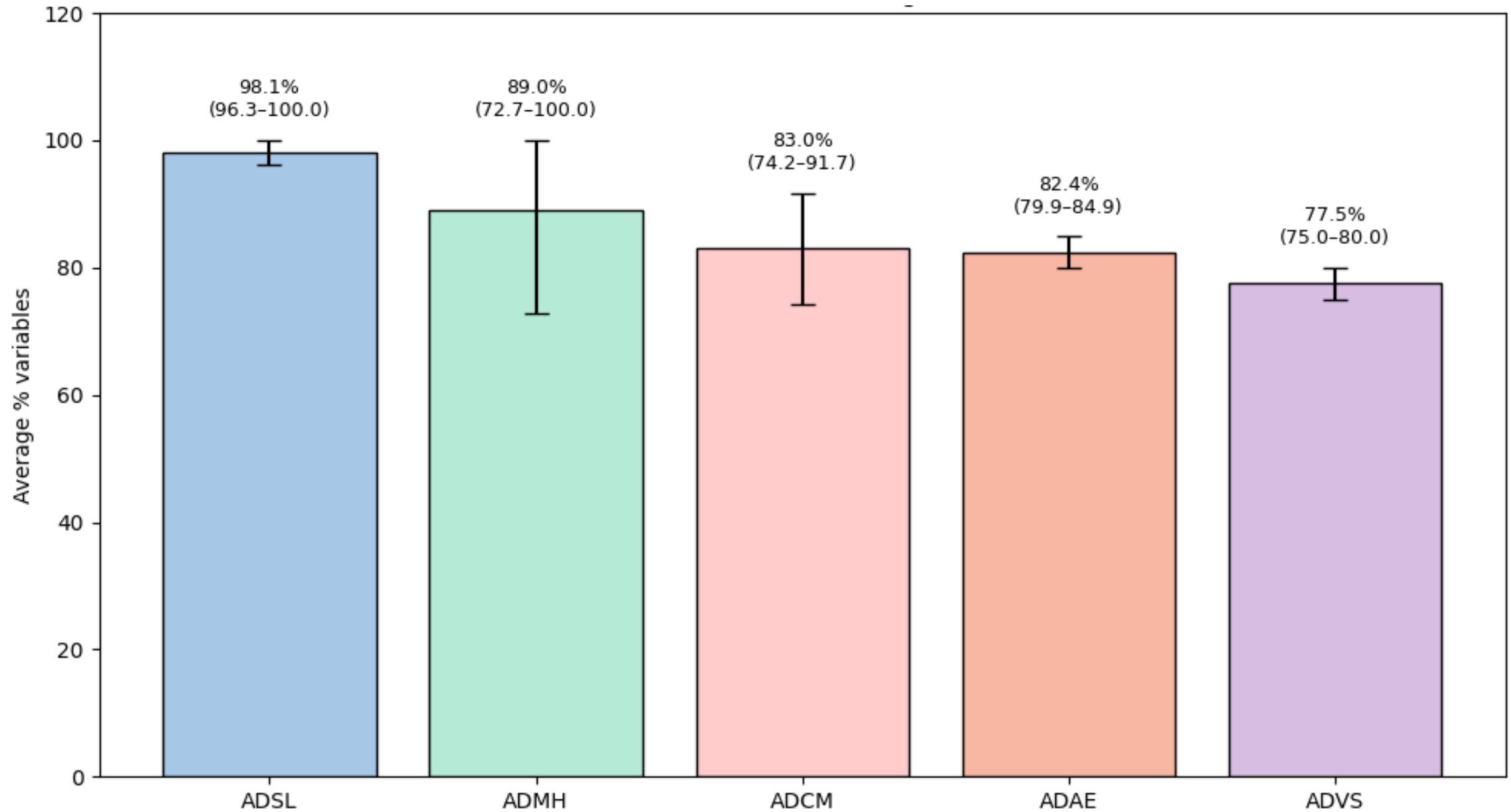


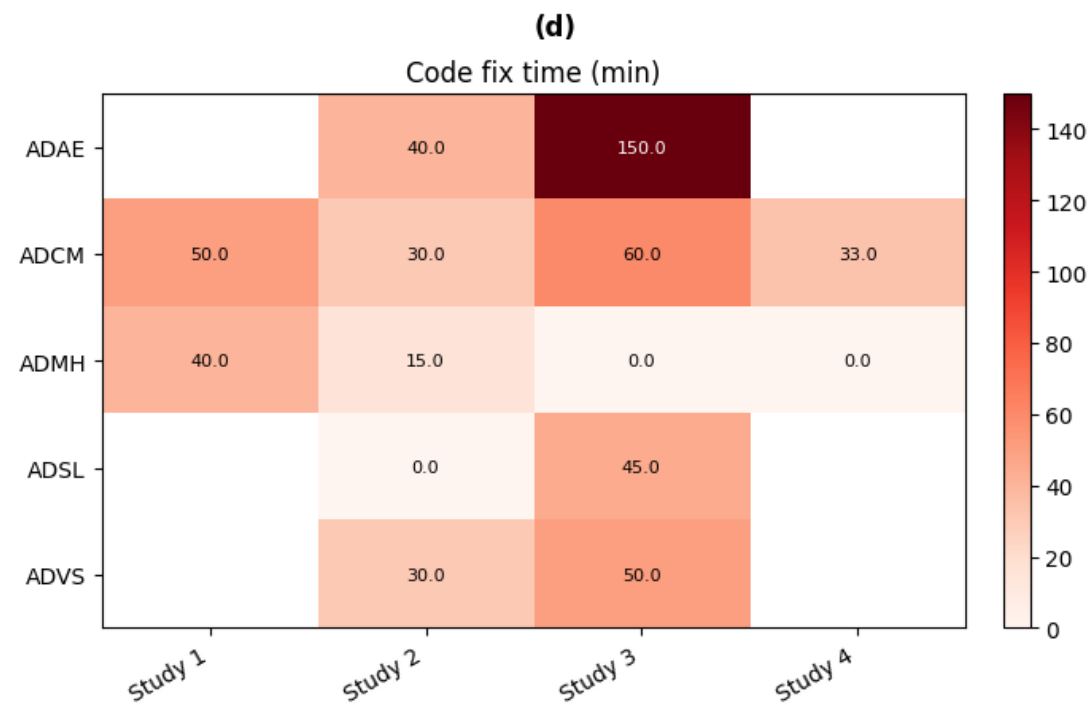
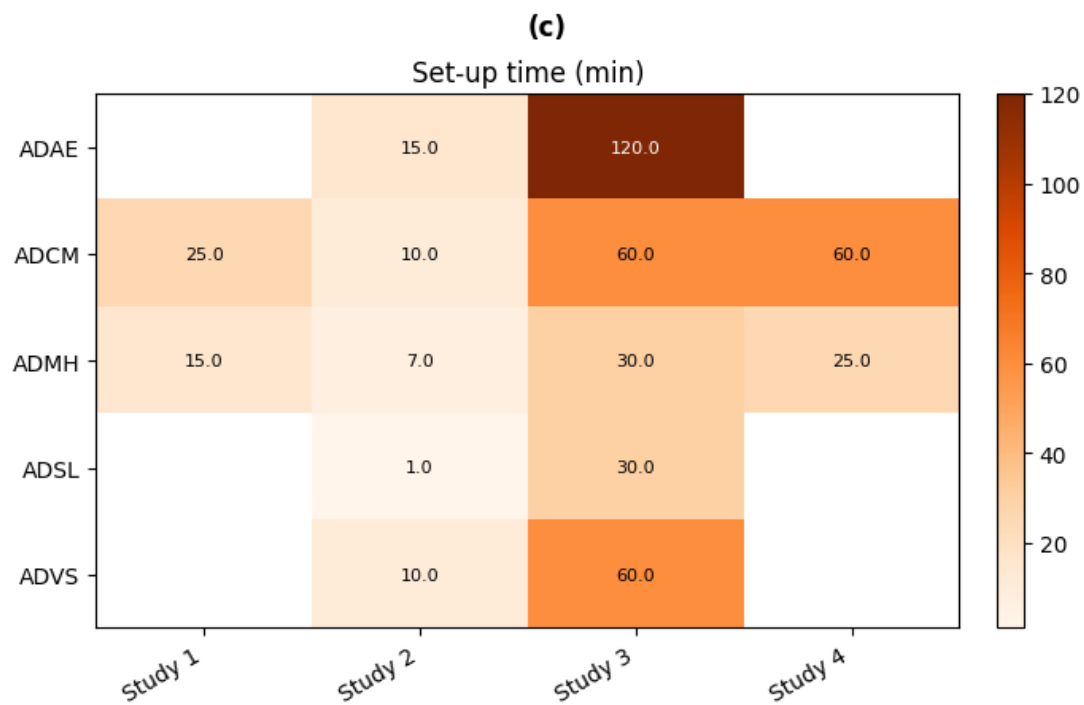
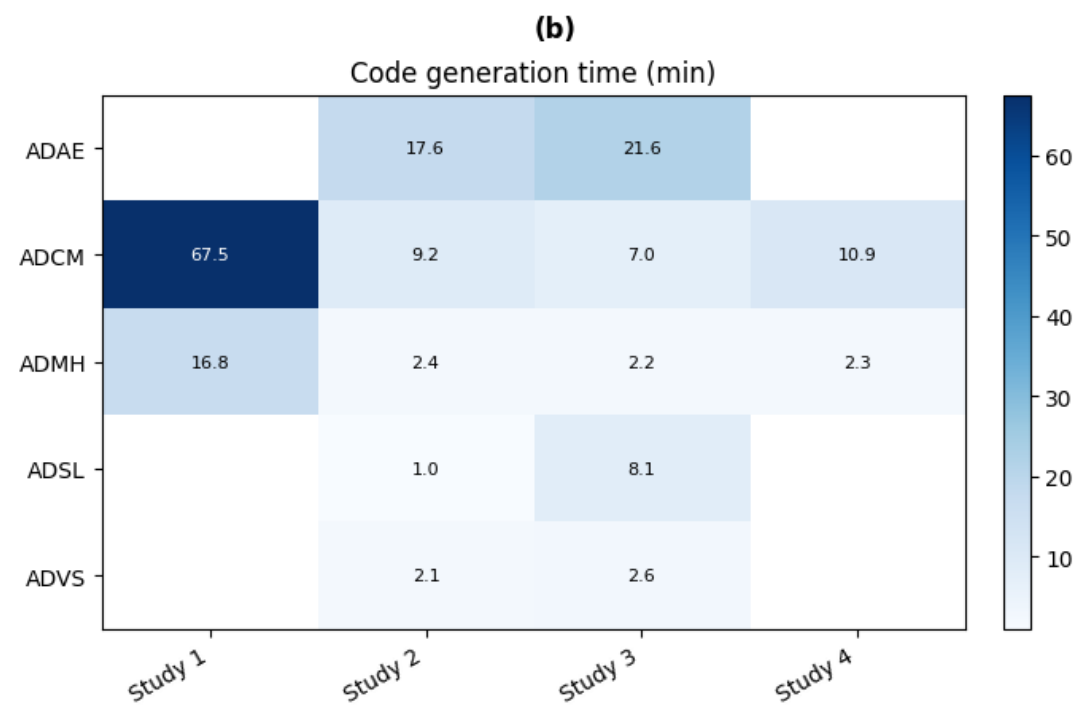
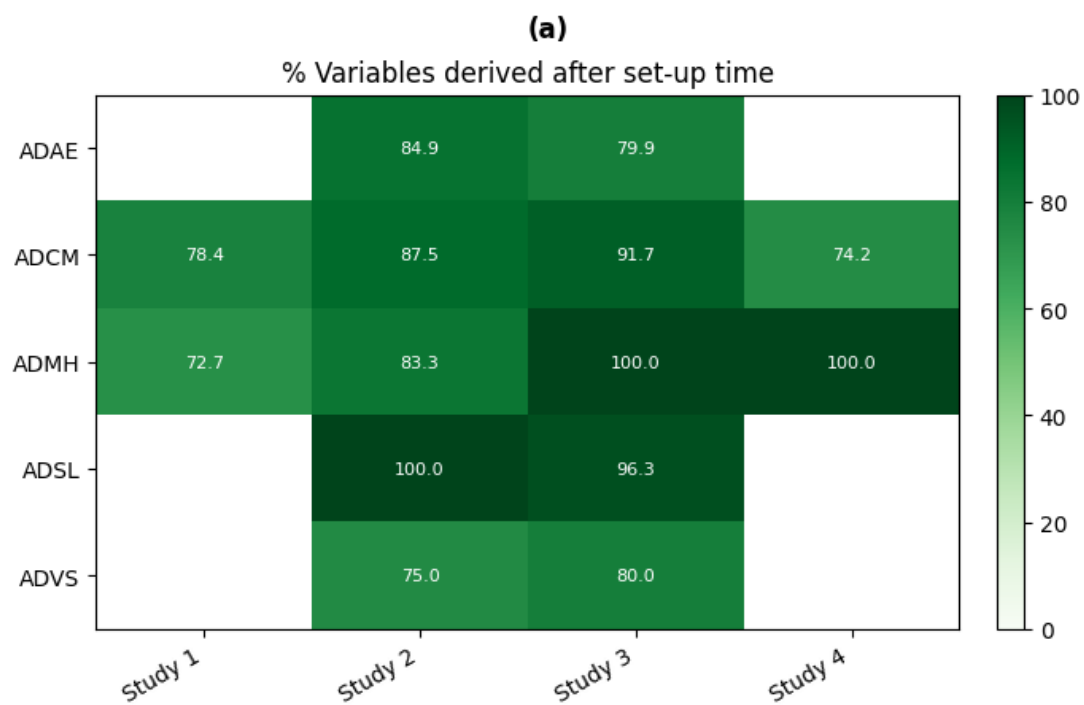
Average End-to-End Time Split per ADaM Dataset

SpectoSAS → Set-Up → Code Fix



Average % variables derived after set up time





Issues Summary

Issues Contributing to Setup

 Libname & Dataset References

 Code Hygiene

 Macro Handling

Issues Contributing to Code Fix Time

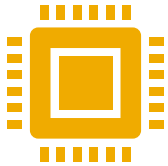
 Derivation Logic Gaps

 Number of Observations Mismatch

 Missing Derivations (Token Limitations)



Benefits



Efficiency:

Faster code generation, allowing to allocate time to other critical tasks.



AI assistance:

Reducing manual programming workload with high success rate.



Streamlined Workflow:

Automates and makes the programming process more efficient.

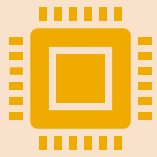


User-Friendly Experience:

Intuitive interface, easy to navigate, accessible for various experience levels.



Recap



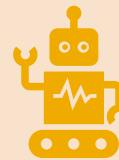
Generates SAS code within 5–15 minutes, with total end-to-end effort reduced to ~1 hour as compared to a few hours traditionally



Automation and AI handle repetitive and complex derivations, while reserving human effort for review and study-specific judgment.



Preserves a human-in-the-loop workflow, ensuring generated code remains reproducible and aligned with regulatory expectations.



Leverages institutional knowledge through reuse, promoting consistency across studies and reducing reimplementation.



Questions?

