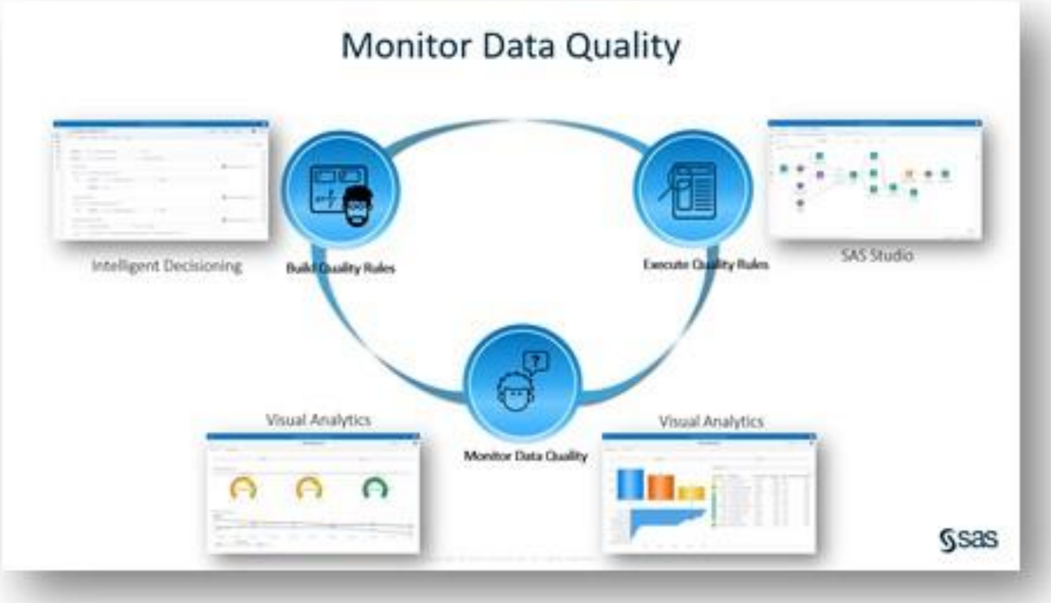
Monitoring Data Quality

> **Note:** This demo is a Proof of Concept. It shows the capabilities for SAS Viya to monitor data quality and to visualize the result in a Data Quality Dashboard.
> For real projects this can used as a starting point for a customer. A customer implementation should consider the specific customer requirements. Especially the flows in SAS Studio as well as the Dashboard itself might need to get adjusted to meet customer requirements.

This document is to help you to better understand the DQ Dashboard demo and to give you enough information to run the demo by yourself.

Below you'll find information about the SAS Viya components used to build the DQ Dashboard as well as some information about what each component is doing in the process.



Monitoring data quality in SAS Viya by using SAS Intelligent Decisioning, SAS Studio and SAS Visual Analytics.

- SAS Intelligent Decisioning (ID) is used to build, test and version business rules for different data quality dimensions.
  ID has a user-friendly UI that allows non-technical users to build business rules with no or little coding.
  Rules can get tested in the same environment to ensure they are fit for purpose.
  Intelligent Decisioning also supports versioning and workflows to govern the business rule life cycle.
- When the business rules are built in Intelligent Decisioning they are invoked in SAS Studio.
  In SAS Studio a Flow job reads data from the data source, for which the data quality should be checked and invokes the business rules from Intelligent Decisioning to check the data quality and writes the result to an output table which Visual Analytics is using to present the data quality dashboard.

- Visual Analytics uses the output from the SAS Studio Flow job to present the data quality dashboard.
  The DQ Dashboard can be used by Data Stewards, and others, to get an understanding of the current data quality state for individual data sources. The user gets an understanding of the data quality for each data source down to field level. The DQ Dashboard also supports Data Stewards to notify data owners to take action if required.

Monitoring Data Quality demo

The Monitoring Data Quality demo is available on ssemonthly.
Note: If the data for the Dashboard is not loaded, please follow the steps in *Prepare Demo* first to load the Dashboard data.

Data Quality Dashboard

The DQ Dashboard is located in: SAS Content/GTPPub/Data Management/DQ Dashboard/Dashboard.

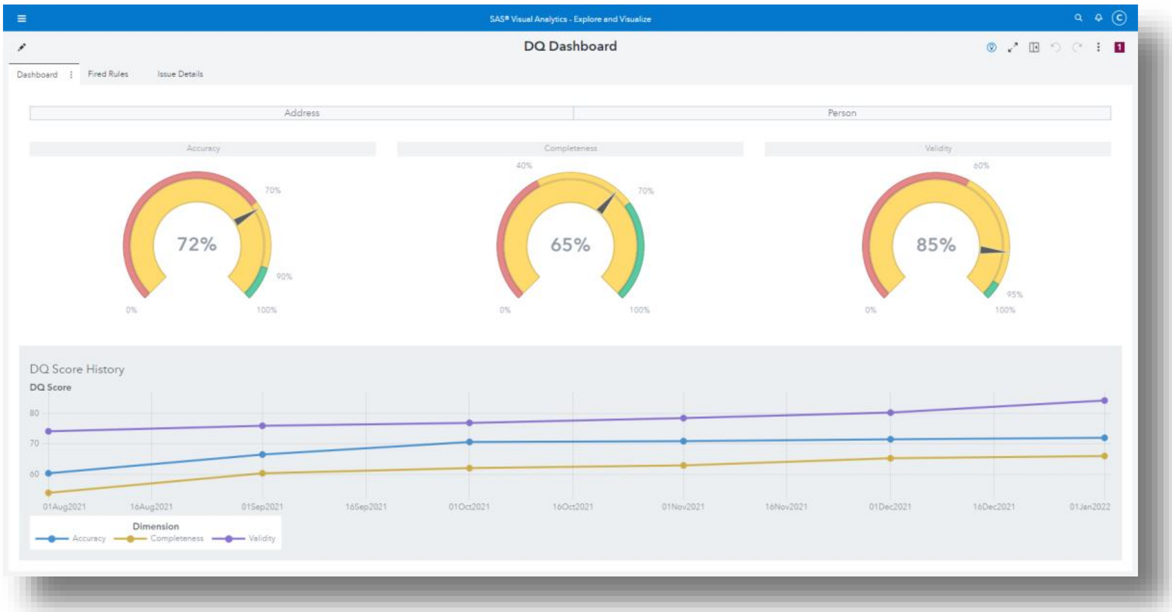
The dashboard is a general DQ Dashboard to visualize the output from data monitoring. It shows some capabilities but could be built more sophisticated depending on the customer requirements.

The dashboard has three tabs ‘Dashboard’, ‘Fired Rules’ and ‘Issue Details’.

Dashboard

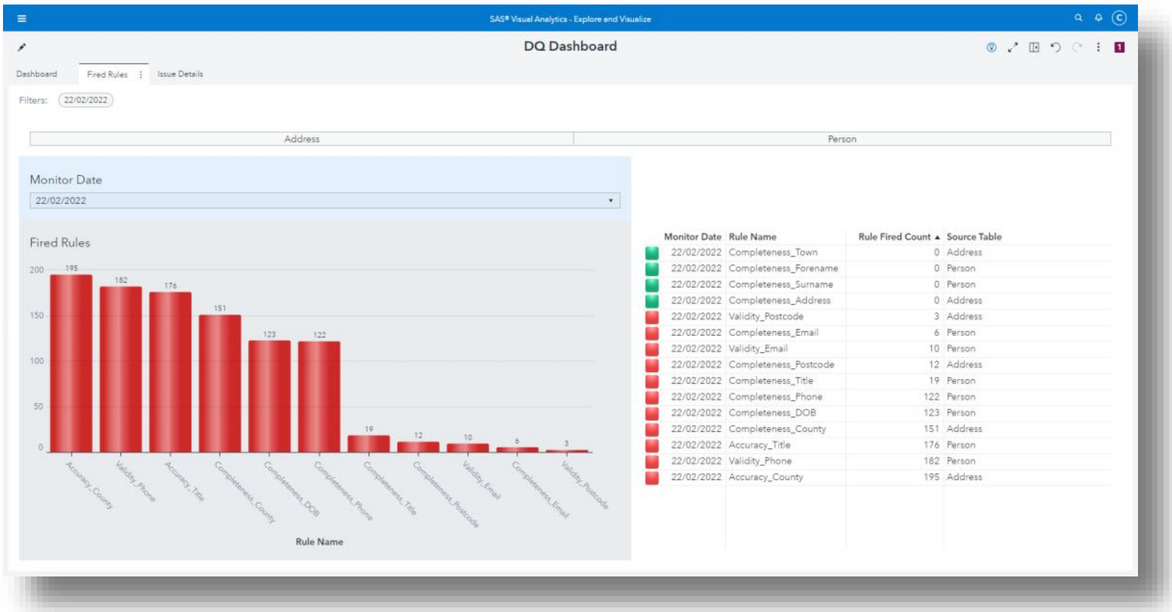
The Dashboard gives a general overview of the overall data quality as well as the data quality for each data source (currently: Address, Person). It shows information for all data quality dimensions that are measured. Currently this is Accuracy, Completeness and Validity but more could be added.

It also shows trending information how the data quality has changed over time for the different dimensions.



Fired Rules

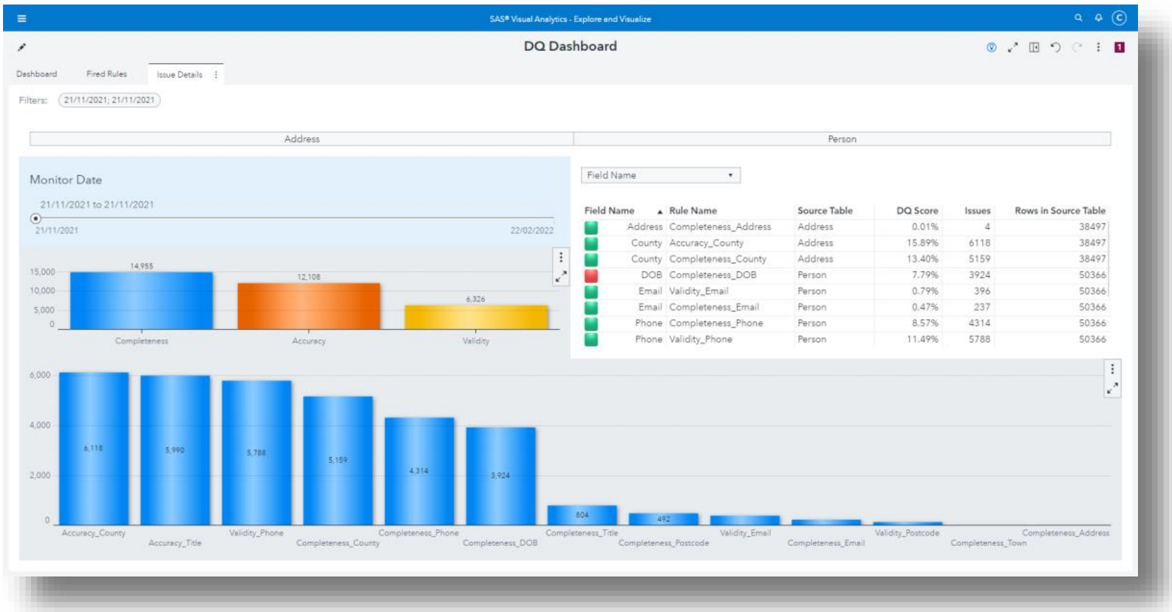
Fired Rules gives an overview how often a Business Rules war triggered at a certain date. The graph shows all fired rules. The list on the right shoes all rules, including the once that were not triggered. If a rule was not fired it is marked green otherwise the rules is marked red.



	Monitor Date	Rule Name	Rule Fired Count ▲	Source Table
🟩	22/02/2022	Completeness_Town	0	Address
🟩	22/02/2022	Completeness_Forename	0	Person
🟩	22/02/2022	Completeness_Surname	0	Person
🟩	22/02/2022	Completeness_Address	0	Address
🟥	22/02/2022	Validity_Postcode	3	Address
🟥	22/02/2022	Completeness_Email	6	Person
🟥	22/02/2022	Validity_Email	10	Person
🟥	22/02/2022	Completeness_Postcode	12	Address
🟥	22/02/2022	Completeness_Title	19	Person
🟥	22/02/2022	Completeness_Phone	122	Person
🟥	22/02/2022	Completeness_DOB	123	Person
🟥	22/02/2022	Completeness_County	151	Address
🟥	22/02/2022	Accuracy_Title	176	Person
🟥	22/02/2022	Validity_Phone	182	Person
🟥	22/02/2022	Accuracy_County	195	Address

Issue Details

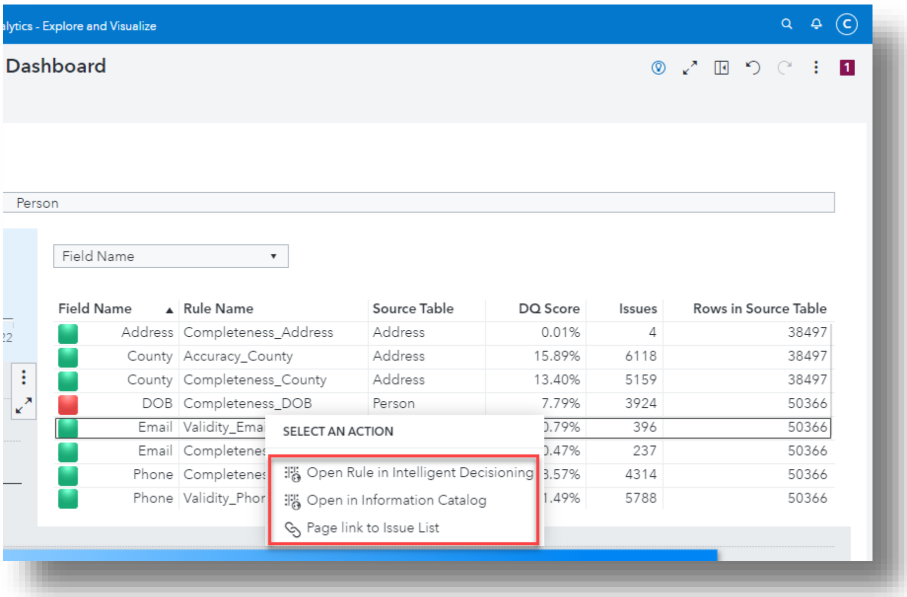
Issue Details give more detailed information on the quality dimension, business rules and checked fields. For the dimensions and business rules it gives information how many issues occurred. For the fields is gives information how many issues occurred as well as the DQ Score and more…



	Field Name ▲	Rule Name	Source Table	DQ Score	Issues	Rows in Source Table
🟩	Address	Completeness_Address	Address	0.01%	4	38497
🟩	County	Accuracy_County	Address	15.89%	6118	38497
🟩	County	Completeness_County	Address	13.40%	5159	38497
🟥	DOB	Completeness_DOB	Person	7.79%	3924	50366
🟩	Email	Validity_Email	Person	0.79%	396	50366
🟩	Email	Completeness_Email	Person	0.47%	237	50366
🟩	Phone	Completeness_Phone	Person	8.57%	4314	50366
🟩	Phone	Validity_Phone	Person	11.49%	5788	50366

When you double click a field line you can:

- open the monitor rule in Intelligent Decisioning by clicking on ‘Open Rules in Intelligent Decisioning’.
- open the monitored table in ‘Information Catalog’.
- see the list of issue records by clicking on ‘Page link to Issue List’.


lytics - Explore and Visualize
Dashboard
Person
Field Name
SELECT AN ACTION
Open Rule in Intelligent Decisioning
Open in Information Catalog
Page link to Issue List

	Field Name	Rule Name	Source Table	DQ Score	Issues	Rows in Source Table
🟩	Address	Completeness_Address	Address	0.01%	4	38497
🟩	County	Accuracy_County	Address	15.89%	6118	38497
🟩	County	Completeness_County	Address	13.40%	5159	38497
🟥	DOB	Completeness_DOB	Person	7.79%	3924	50366
🟩	Email	Validity_Emai		).79%	396	50366
🟩	Email	Completenes		).47%	237	50366
🟩	Phone	Completenes		3.57%	4314	50366
🟩	Phone	Validity_Phor		1.49%	5788	50366

Data Quality Rules in Intelligent Decisioning

In Intelligent Decisioning the Monitoring Rules are built in three steps:

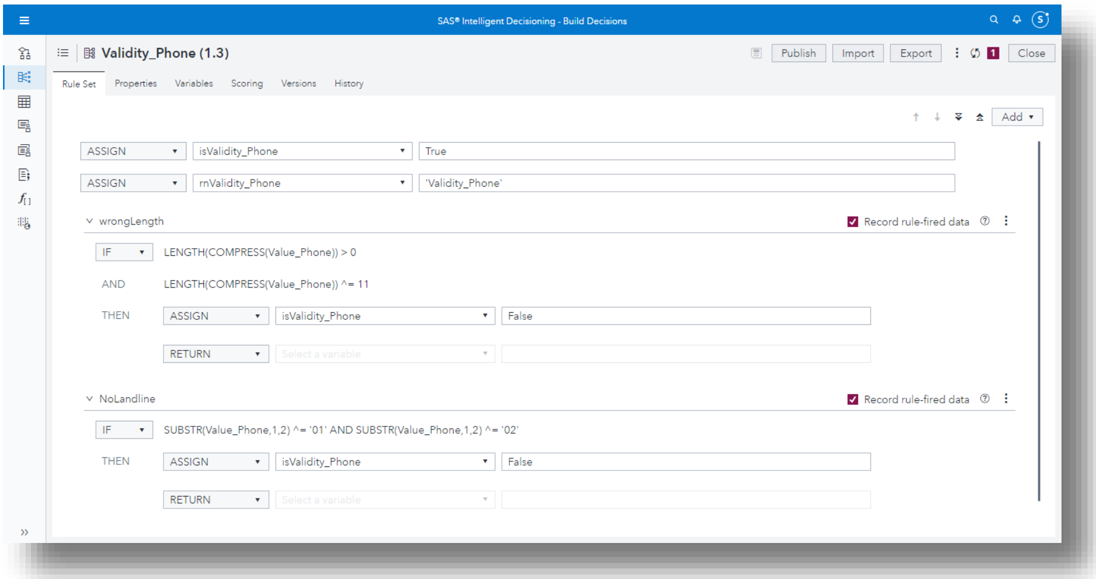
- DQ Business Rules
- DQ Fields
- Monitoring Tasks

DQ Business Rules

The Business Rules are located in: SAS Content/GTPPub/Data Management/DQ Dashboard/Monitor Data Quality/Rules.

DQ Business Rules describe the data quality rules for the different quality dimensions. The rules are implemented as Rules Sets in Intelligent Decisioning. If possible, the rules are written to be reusable, so that a rule can be used for different fields. For example, rule “Completeness String” is used to check completeness on most character columns.

Rules can be built with little or no coding by “clicking” If-Then_Else statements together.



ASSIGN isValidity_Phone True

ASSIGN rnValidity_Phone 'Validity_Phone'

wrongLength

IF LENGTH(COMPRESS(Value_Phone)) > 0

AND LENGTH(COMPRESS(Value_Phone)) ^= 11

THEN ASSIGN isValidity_Phone False

RETURN Select a variable

NoLandline

IF SUBSTR(Value_Phone,1,2) ^= '01' AND SUBSTR(Value_Phone,1,2) ^= '02'

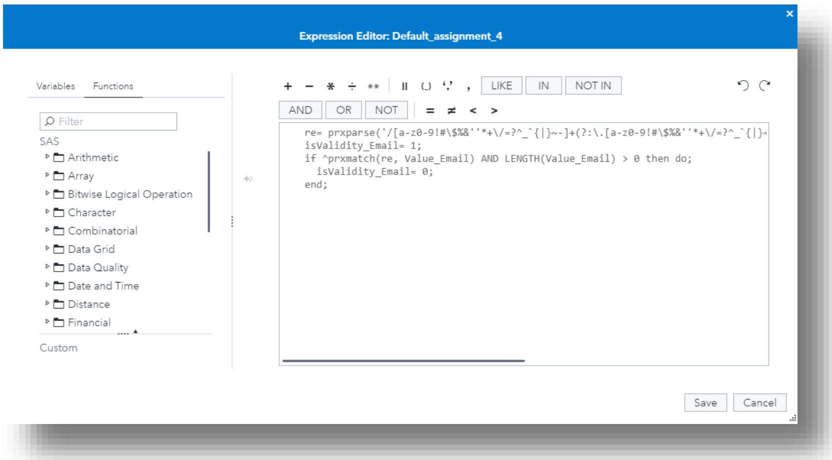
THEN ASSIGN isValidity_Phone False

RETURN Select a variable

For more sophisticated rules Intelligent Decisioning Lookup Tables could also be used. For example, ‘Accuracy_Title’.

It is also possible to code rules if necessary. When coding rules this is done in SAS DS2 language. For this a rich set of DS2 functions is available including functions to call the QKB.

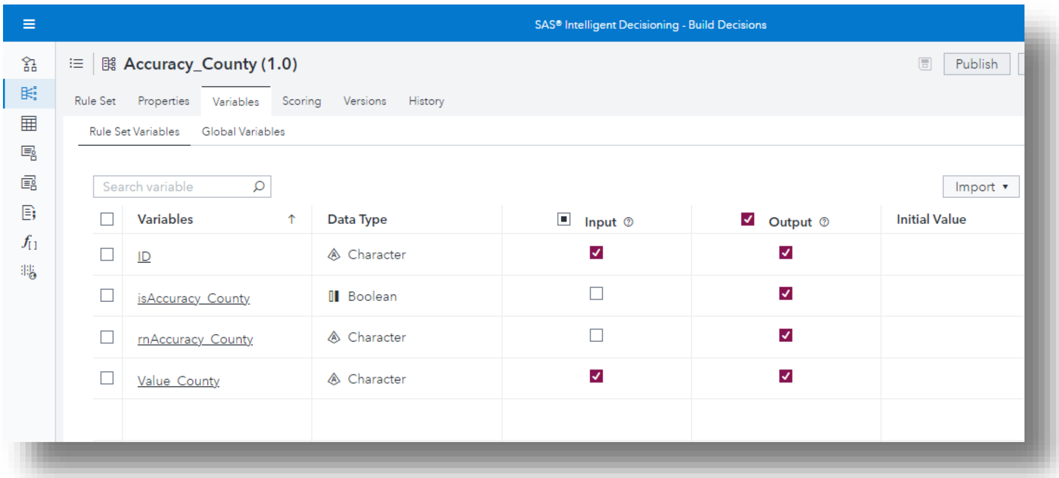
You could also build your own DS2 functions in Intelligent Decisioning and use them in a reusable manner in Rules Sets if needed.



```
re= prxparse('/[a-z0-9!#\$%&''*+\/=?^_`{|}~-]+(?:\.[a-z0-9!#\$%&''*+\/=?^_`{|}~
isValidity_Email= 1;
if ^prxmatch(re, Value_Email) AND LENGTH(Value_Email) > 0 then do;
  isValidity_Email= 0;
end;
```

To be able to show the monitoring result in the DQ Dashboard, all rules must follow the same pattern and naming conventions:

- All rules have as input parameters
  - the record ID
    - Field Name: **ID**
  - the value to be checked. The value to be checked could be spread over one or more fields (see rules for address).
    - Field Name: **value_<field name>** e.g.: value_County
- The output parameters are the same as the input parameters plus two additional fields.
  - one field to indicate if the rule was triggered
    - Field Name: **is<DQ Dimension>_<field name>** e.g.: isAccuracy_County
  - one field to carry the rules name, which is needed as information field in the DQ Dashboard in VA.
    - Field Name: **rn<DQ Dimension>_<field name>** e.g.: rnAccuracy_County



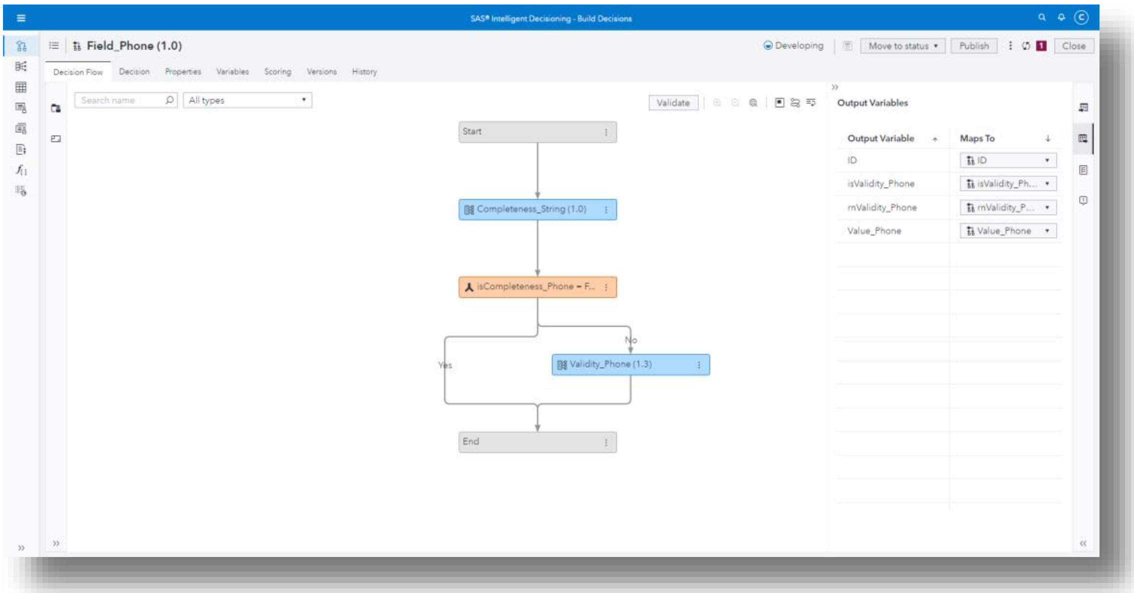
□	Variables ↑	Data Type	▣ Input	☑ Output	Initial Value
□	ID	Character	☑	☑	
□	isAccuracy_County	Boolean	□	☑	
□	rnAccuracy_County	Character	□	☑	
□	Value_County	Character	☑	☑	

When a rule is created, it can get tested within Intelligent Decisioning to ensure the rules works as desired before calling in through SAS Studio.

DQ Fields

DQ Fields are located in: SAS Content/GTPPub/Data Management/DQ Dashboard/Monitor Data Quality/Fields.

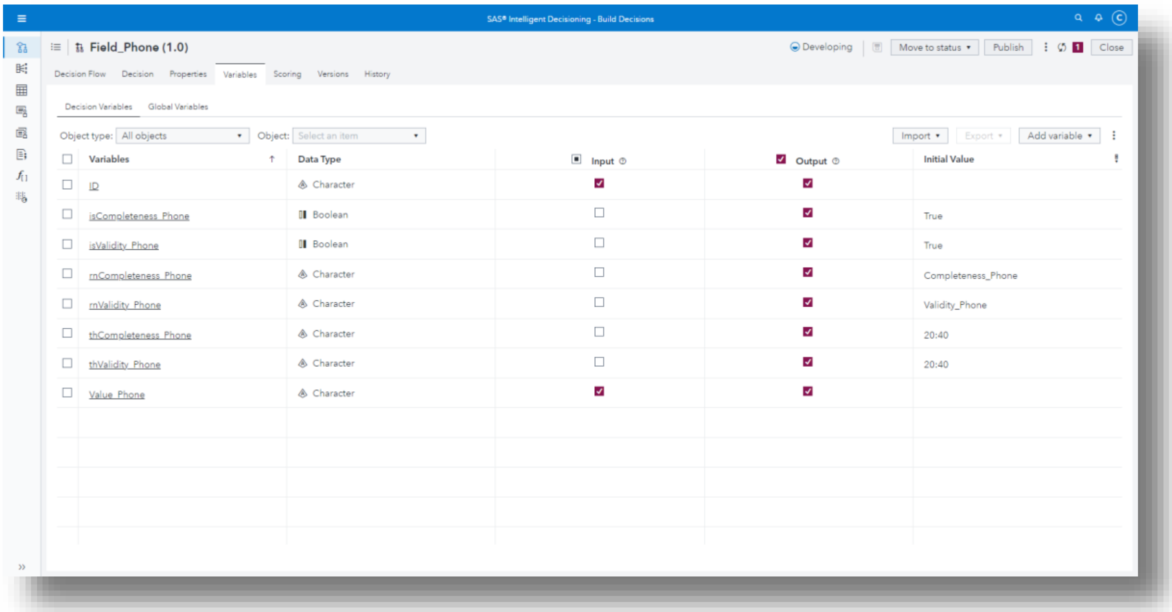
DQ Fields are decision flows in Intelligent Decisioning. These decision flows describe all Business Rules that belong to one monitoring field.
In the decision flow a rule for one quality dimension is checked, if the rule was fired no other rules will be checked. If the rule was not fired the next quality dimension rule will be checked, and so on…



Output Variable	Maps To
ID	ID
isValidity_Phone	isValidity_Ph...
rnValidity_Phone	rnValidity_P...
Value_Phone	Value_Phone

At DQ Fields level we also set the threshold for a rule. The threshold describes when a rule is marked green, yellow or red in the Dashboard in tab ‘Issue Details’.

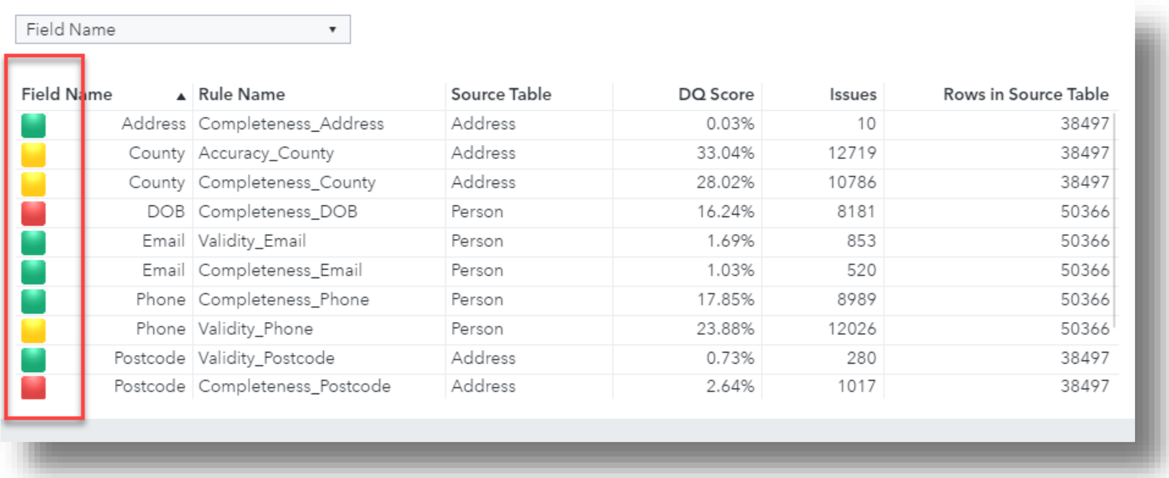
- Field Name: **th<DQ Dimension>_<field name>** e.g.: thAccuracy_County



Object type: All objects Object: Select an item

□	Variables	Data Type	Input	Output	Initial Value
□	ID	Character	☑	☑	
□	isCompleteness_Phone	Boolean	□	☑	True
□	isValidity_Phone	Boolean	□	☑	True
□	rnCompleteness_Phone	Character	□	☑	Completeness_Phone
□	rnValidity_Phone	Character	□	☑	Validity_Phone
□	thCompleteness_Phone	Character	□	☑	20:40
□	thValidity_Phone	Character	□	☑	20:40
□	Value_Phone	Character	☑	☑	

For example, if the field is set to 20:40 then a score between 0 and 20 will mark the rule green, a score between 21 and 40 will mark the rule yellow and a score over 40 will mark the rule red.



Field Name ▾

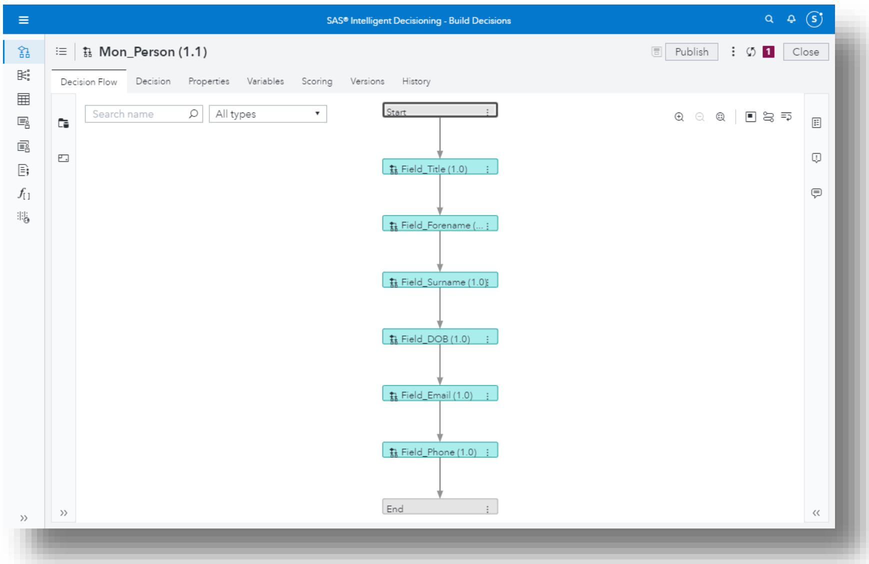
Field Name ▲		Rule Name	Source Table	DQ Score	Issues	Rows in Source Table
🟩	Address	Completeness_Address	Address	0.03%	10	38497
🟨	County	Accuracy_County	Address	33.04%	12719	38497
🟨	County	Completeness_County	Address	28.02%	10786	38497
🟥	DOB	Completeness_DOB	Person	16.24%	8181	50366
🟩	Email	Validity_Email	Person	1.69%	853	50366
🟩	Email	Completeness_Email	Person	1.03%	520	50366
🟩	Phone	Completeness_Phone	Person	17.85%	8989	50366
🟨	Phone	Validity_Phone	Person	23.88%	12026	50366
🟩	Postcode	Validity_Postcode	Address	0.73%	280	38497
🟥	Postcode	Completeness_Postcode	Address	2.64%	1017	38497

On this level you can also run tests in Intelligent Decisioning to make sure all rules on a monitoring field work as desired.

Monitoring Task

The Monitoring Tasks are located in: SAS Content/GTPPub/Data Management/DQ Dashboard/Monitor Data Quality/Monitor Tasks.

The Monitoring Tasks are decision flows where all DQ Field rules are called that belong to one source (table) to be monitored. These decision flows will be called to monitor a data source.



For example, for data source Person we are going to check on columns Title, Forename, Surname, DOB, Email and Phone as shown in the screenshot above.

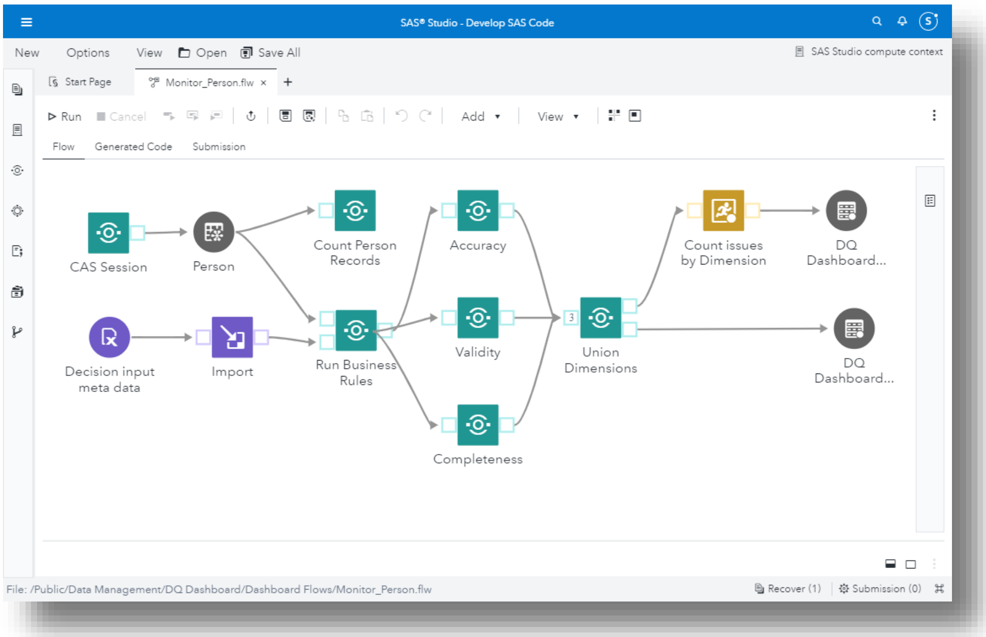
Also, at this level the decision can get tested in Intelligent Decisioning before executing it in SAS Studio.

Executing DQ Monitoring in SAS Studio

The monitoring tasks are executing through Studio Flow jobs.
The jobs are located in: SAS Content/GTPPub/Data Management/DQ Dashboard/Dashboard Flows.

For each data source to be monitored we have one Studio Flow: Monitor_Person.flw and Monitor_Address.flw.

These Studio Flows use a set on custom Steps to support the monitoring process.

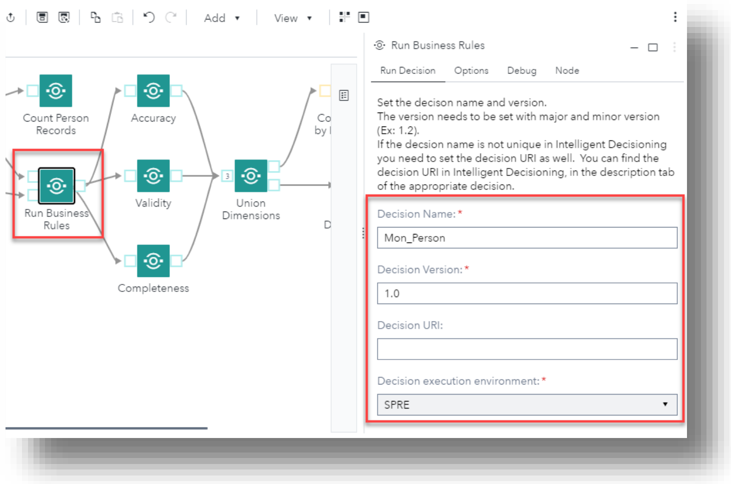


In the Studio Flow we point at the data source that we are going to monitor. For example, Person.

From Person we call the Step ‘Run Business Rules’. This custom step calls the decision flow for version 1.0 of Monitoring Task Mon_Person. Mon_Person wil be executed in SPRE.

Behind this custom step we call the decision flow via the ID macro %DCM_EXECUTE_DECISION which allows us to choose the execution environment for the decision flow to be SPRE or CAS.

In the current version of Studio Engineer we could use step *Execute Decisions* instead of a customer step. However, using the step *Execute Decisions* would prevent us from using a Query node in ID which we might have to use for DQ Dimension Uniqueness. We also could not control where we run the decision flow (CAS or SPRE) as step *Execute Decisions* will run the decision flow in CAS.



Set the decison name and version.
The version needs to be set with major and minor version (Ex: 1.2).
If the decsion name is not unique in Intelligent Decisioning you need to set the decision URI as well. You can find the decision URI in Intelligent Decisioning, in the description tab of the appropriate decision.

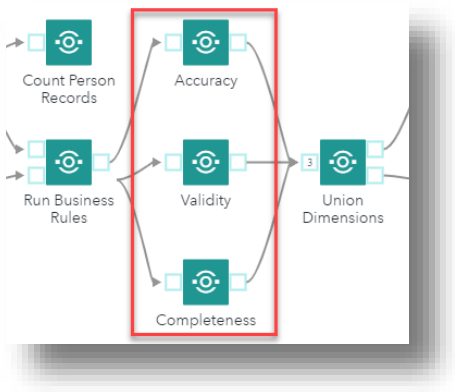
Decision Name: * Mon_Person

Decision Version: * 1.0

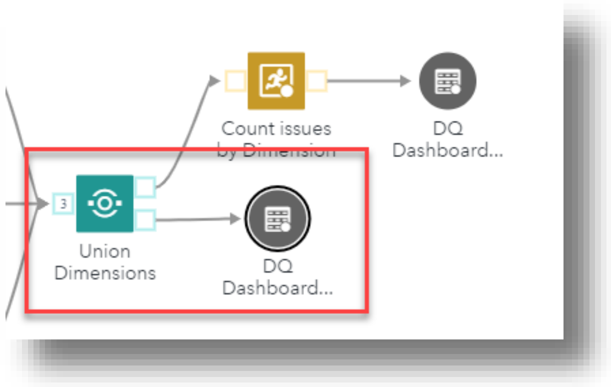
Decision URI:

Decision execution environment: * SPRE

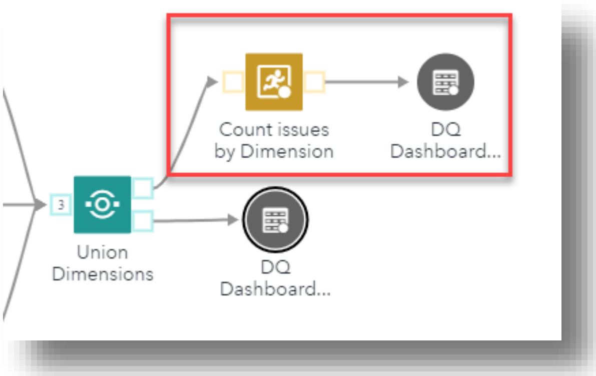
After Run Business Rules we split the output into the DQ Dimensions (Accuracy, Validity, Completeness) to get information for each dimension as needed in the DQ Dashboard.



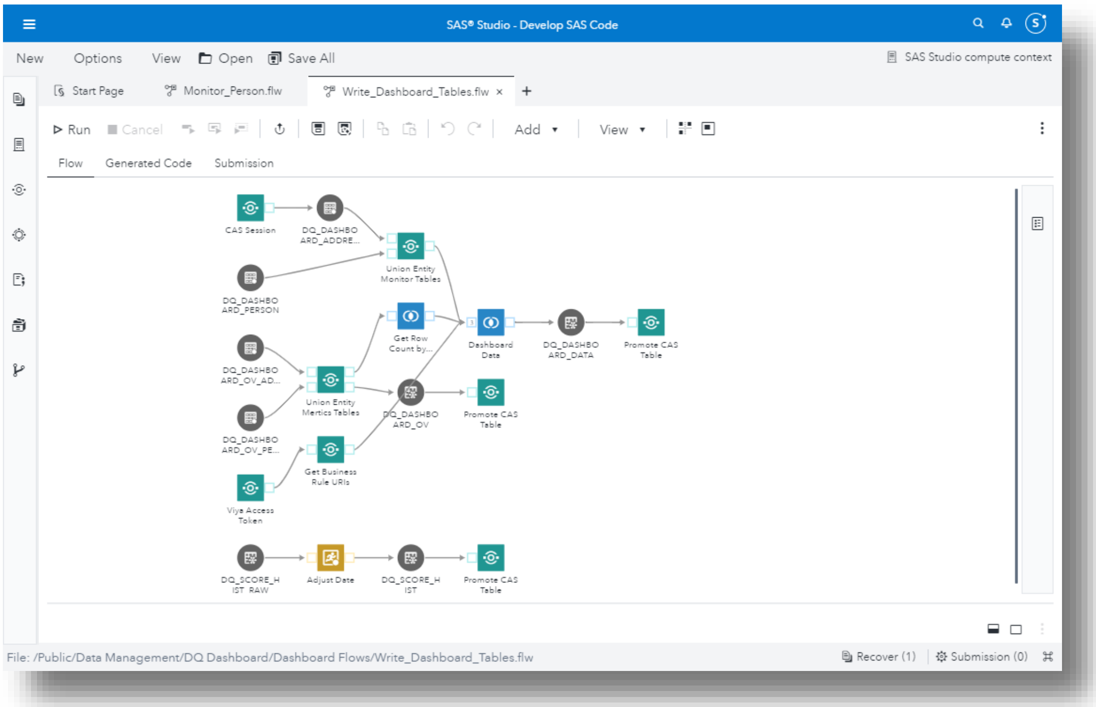
In the next step (Union Dimensions) we combine all outputs from the dimensions split steps into one table and output the result into the final flow table.



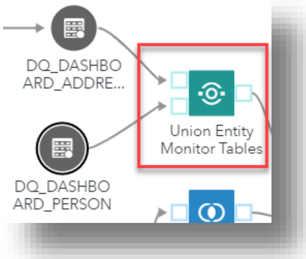
In another branch from Union Dimensions, we write a second flow output table to store monitor metrics for the data source required by DQ Dashboard.

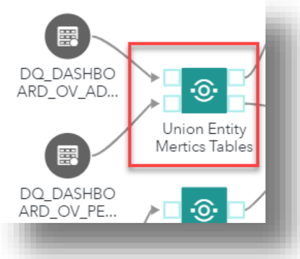


To create the data for DQ Dashboard we run the Studio flows for Monitor_Person and Monitor_Address. After this we run the Studio flow Write_Dashboard_Tables.flw to combine the output from the monitoring flows and to prepare the data needed in the Dashboard.

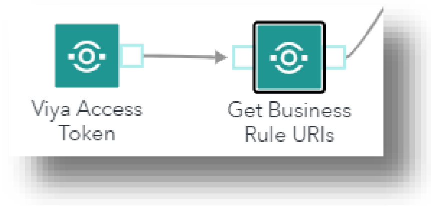


Each monitoring flow has two output tables. Write_Dashboard_Tables flow combines the tables of the same type.

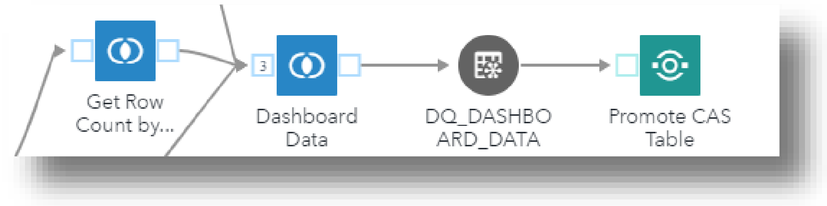




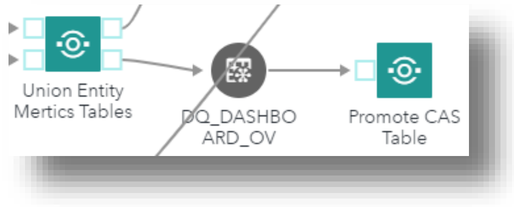
We also generate deep link information for the dashboard to open a Business Rule in Intelligent Decisioning from DQ Dashboard.



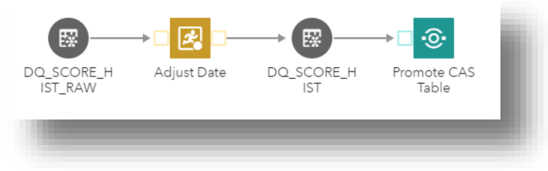
We add some more information to the Dashboard Data table, output the table and promote the table so we can use it in VA for the DQ Dashboard.



From the Step 'Union Entity Metrics Tables' we output the metrics table for the dashboard and promote it to use it in the Dashboard.



For the DQ trending information in the Dashboard we generate some demo dummy data in the flow and just update the timestamp information to make the timing information in the Dashboard more current.

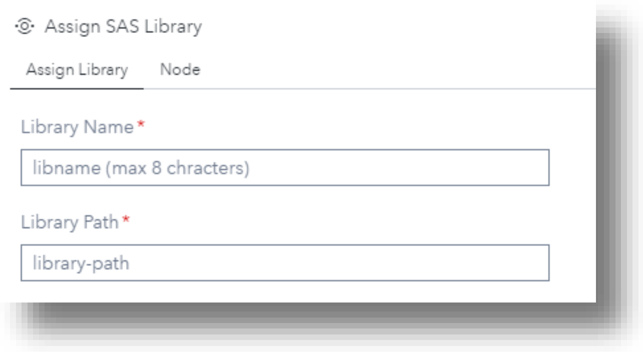


Custom Steps

Here are all custom steps described that we have used in SAS Studio to monitor the data and to prepare the output for the DQ Dashboard.

Assign SAS Library

This step assigns a SAS library.



Assign SAS Library

Library Name *

libname (max 8 chracters)

Library Path *

library-path

Library Name:

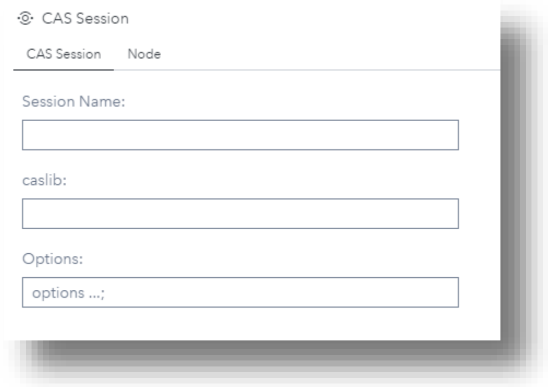
Set the library name you want to assign.

Library Path:

Point to the library path on the hard drive where to store the dataset file.

CAS Session

This step creates a CAS session.



CAS Session

Session Name:

caslib:

Options:

options ...;

Session Name:

Set the CAS session name. If no session name is supplied a default session name (mySession) will be used.

Caslib:

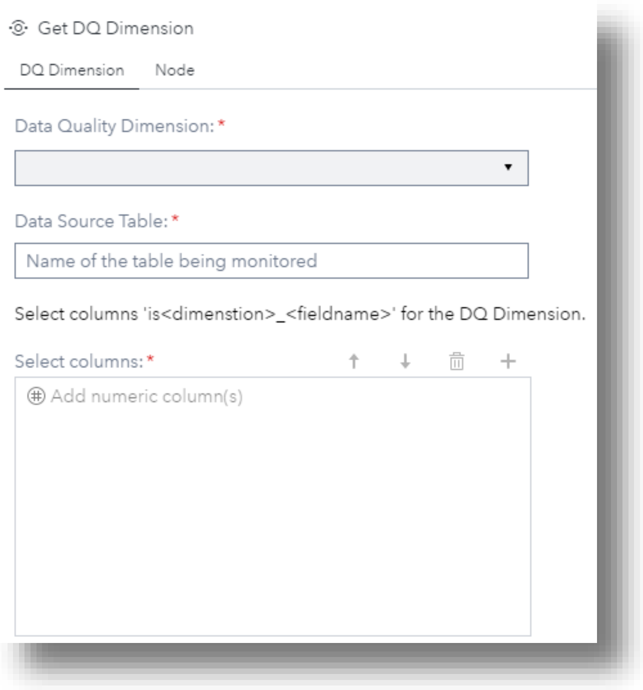
Set the caslib for the session. For example, CASUSER.

Options:

Set CAS session options.

Get DQ Dimension

This Step is called after a monitor rule was executed. It groups all return values for one DQ dimension of a monitor flow together and moves the result to a table in transposed form.


Get DQ Dimension

DQ Dimension | Node

Data Quality Dimension: *

Data Source Table: *

Name of the table being monitored

Select columns 'is<dimenstion>_<fieldname>' for the DQ Dimension.

Select columns: *

Add numeric column(s)

Data Quality Dimension:

Select the dimension to group on from the drop-down list.

Data Source Table:

Set the source table name. This information is required by the dashboard to state where a record originated from.

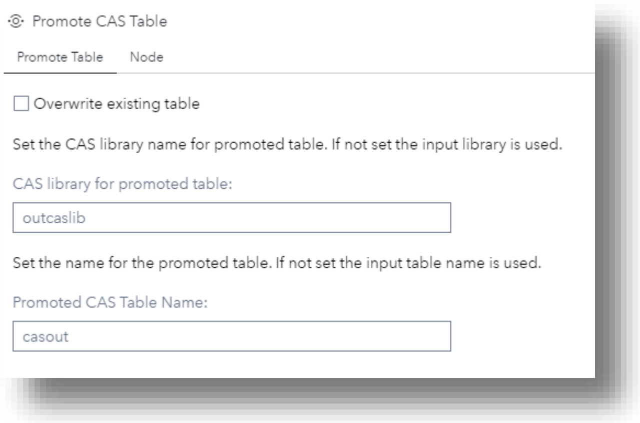
Select Columns:

Select the return columns from the monitoring flow you want to filter on with this step.

Each monitoring rule returns a parameter to indicate if a DQ rule was triggered. The parameter has the naming pattern is<dimension>_<fieldname>. For example, isCompleteness_Phone is the parameter to state the result for field Phone of the quality dimension Completeness.

Promote CAS Table

This Step promotes a table in CAS.


Promote CAS Table

Promote Table | Node

□ Overwrite existing table

Set the CAS library name for promoted table. If not set the input library is used.

CAS library for promoted table:

outcaslib

Set the name for the promoted table. If not set the input table name is used.

Promoted CAS Table Name:

casout

Overwrite existing table:

Tick the box if the step should overwrite a table with the same name in the output library.

CAS library for promoted table:

Name the CAS library for the promoted CAS table. If no library is set the table will be promoted in the library it resides in.

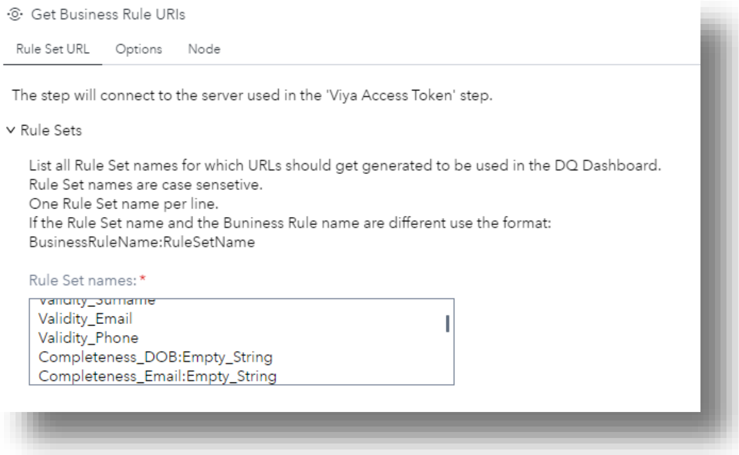
Promoted CAS Table Name:

The name for the promoted table. If it is not set the source table name and the promoted table name are the same.

Rule Set URI

This step generates URIs for Rule Sets in Intelligent Decisioning. These URIs can be used in the Dashboard as a deep link to open a DQ Rule in Intelligent Decisioning from the DQ Dashboard.

This step requires a Viya Access Token. The token is supplied by step ‘Viya Access Token’ and needs to be called before step ‘Rule Set URI’.



The step will connect to the server used in the 'Viya Access Token' step.

Rule Sets

List all Rule Set names for which URLs should get generated to be used in the DQ Dashboard.
Rule Set names are case sensetive.
One Rule Set name per line.
If the Rule Set name and the Buniness Rule name are different use the format:
BusinessRuleName:RuleSetName

Rule Set names: *

```
Validity_Surname
Validity_Email
Validity_Phone
Completeness_DOB:Empty_String
Completeness_Email:Empty_String
```

Rule Set names:

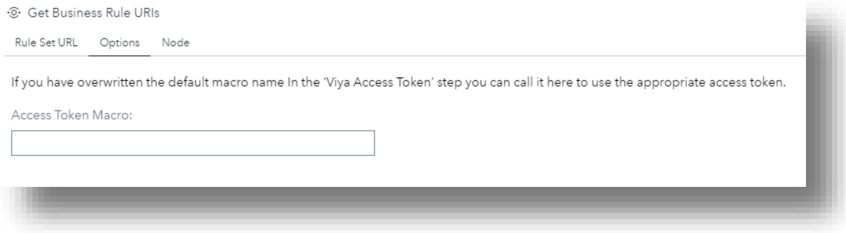
List all Intelligent Decisioning Rule Set names for which a URI should get generated.

Add one Rule Set name per line in the edit box. Note that Rule Set name are case sensitive.

If you have reusable DQ rules, the Rule Set name and the Business rule name may differ. For example, Rule Set name Empty_String is used to check Completeness on character fields. In this case you state both the Business Rule name and Rule Set name separated by a colon like, Completeness_Email: Empty_String

Access Token Marco

Under tab options you can set the macro name to use if the token macro name differs from the default name. See also step ‘Viya Access Token’.



If you have overwritten the default macro name In the 'Viya Access Token' step you can call it here to use the appropriate access token.

Access Token Macro:

Run Decision

This step runs a decision flow from Intelligent Decisioning.

To execute the decision flow the step uses the ID Studio Macro %DCM_EXECUTE_DECISION.

This step has two input ports.

Port inTable1:

Connect the data source table for the decision flow to this port.

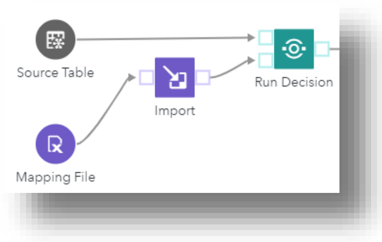
Port inMetaTable1:

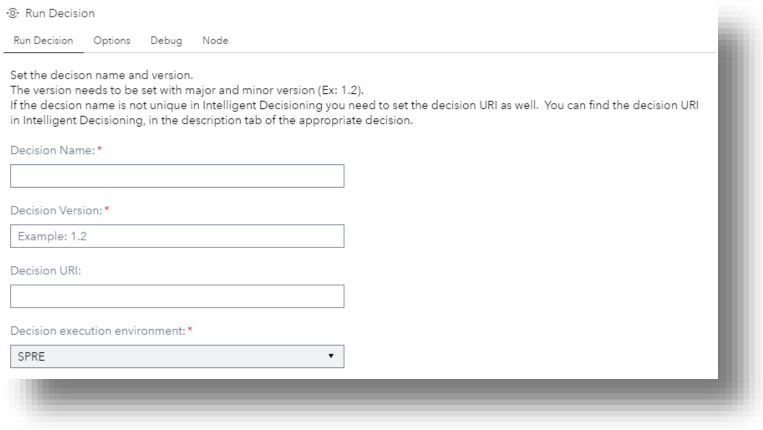
Connect the mapping information to this port. The mapping information is an Excel file to map the data source column name to the decision flow input parameter name. You can map the table column name to the Decision input parameter name and set the data type if necessary.

The Excel input file looks like:

TableColumnName	DecisionParameterName	DataType
City	Value_Town	$50.
Postcode	Value_Postcode	

In a flow you use the file input step to connect to the mapping file, use an Import step to read the file and then connect to the inMetatable1 port.





Set the decison name and version.
The version needs to be set with major and minor version (Ex: 1.2).
If the decsion name is not unique in Intelligent Decisioning you need to set the decision URI as well. You can find the decision URI in Intelligent Decisioning, in the description tab of the appropriate decision.

Decision Name: *

Decision Version: *
Example: 1.2

Decision URI:

Decision execution environment: *
SPRE ▾

Decision Name:

Set the name of the decision from Intelligent Decisioning. Decision names are case sensitive.

Decision Version:

Set the version number of the decision flow in Intelligent Decisioning to be executed.

Decision URI:

If a decision name in Intelligent Decisioning in not unique, you need to set the decision URI. You get the decision URI by opening the decision in Intelligent Decisioning and copy the Object URI from the Properties tab.

Decision execution environment:

Select from the drop-down list where the decision flow should be executed, in SPRE or CAS.

For all other input parameters please see the %DCM_EXECUTE_DECISION macro documentation.

Table Count

This step counts all records in a table and passes the value to a macro. The macro can then be used at a later step in the flow.

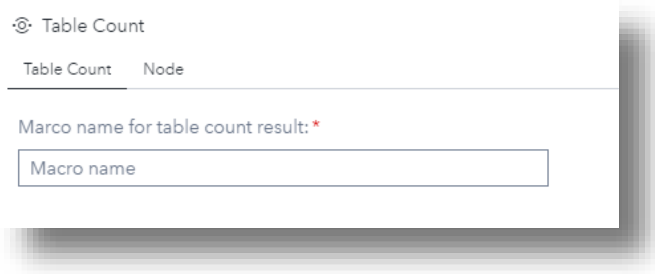


Table Count

Table Count | Node

Marco name for table count result: *

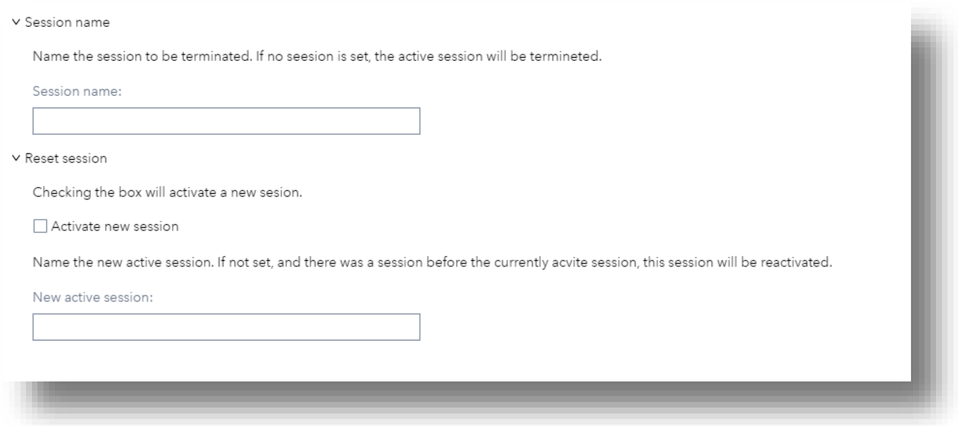
Macro name

Macro name for table count result:

Set the macro name for the table the count result.

Terminate CAS session:

This step is to terminate a CAS session.



Session name

Name the session to be terminated. If no seesion is set, the active session will be termineted.

Session name:

Reset session

Checking the box will activate a new sesion.

□ Activate new session

Name the new active session. If not set, and there was a session before the currently acvite session, this session will be reactivated.

New active session:

Session name:

Set the CAS session to be terminated. If no session is set the active session will be terminated.

Activate new session:

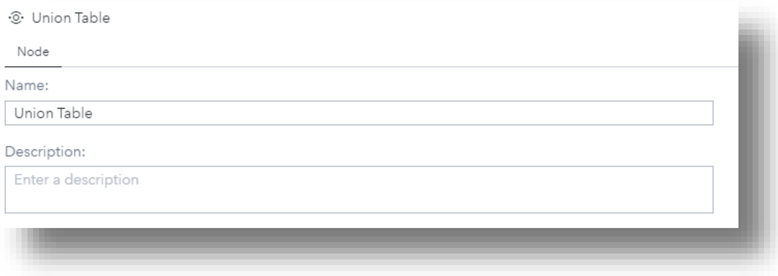
Tick the box if a new session should be activated after the session has been terminated.

New active session:

Set the name for the new CAS session. If no name and there was an active session before the just terminated session, this session will be reactivated.

Union Table

This step unions two or more tables if with the same structure.



Union Table

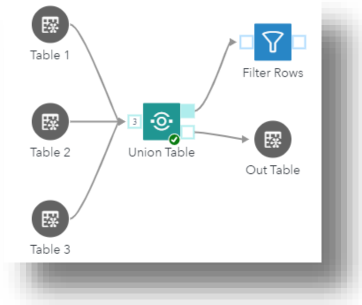
Name:

Union Table

Description:

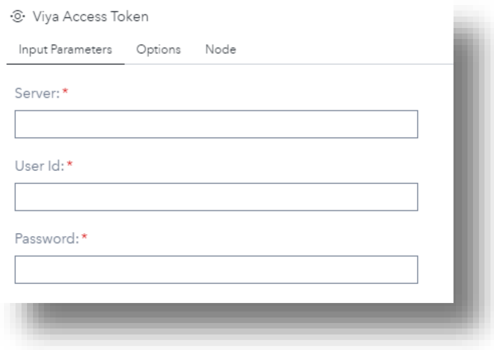
Enter a description

The step supports several output ports. These can be used if you want to go after this step to a table output step as well as to another step. For this scenario you need two output ports.



Viya Access Token

This step retrieves a Viya access token that can be used with steps that call a REST API into Viya.



Viya Access Token

Server: *

User Id: *

Password: *

Server:

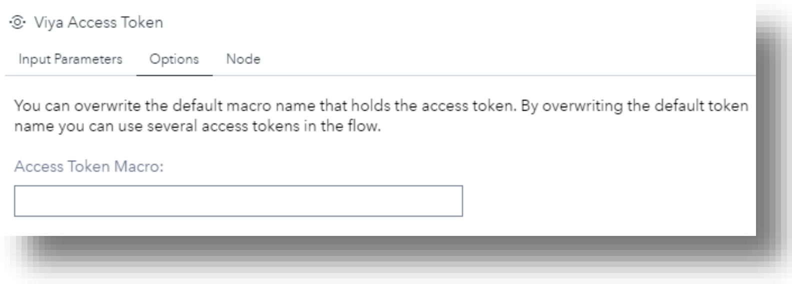
Set the IP address or server name for Viya

User Id:

Set the Viya user id for which to retrieve the access token.

Password:

Set the password for the Viya user id.



You can overwrite the default macro name that holds the access token. By overwriting the default token name you can use several access tokens in the flow.

Access Token Macro:

Prepare Demo on ssemonthly

To ensure that DQ Dashboard demo works a desired follow the steps below to load the data into memory and to run the Studio Flows to load the DQ Dashboard tables.

1. Go to SAS Studio (Develop SAS Code)
    1.1. Go to location: SAS Content/GTPPub/Data Management/DQ Dashboard/SAS Programs
    1.2. Open and run SAS job: Load tables into memory.sas
    1.3. Go to location: SAS Content/ GTPPub /Data Management/DQ Dashboard/Dashboard Flows
    1.4. Open and run flows:
    Monitor_Person.flw
    Monitor_Address.flw
    Write_Dashboard_Tables.flw
2. Go to SAS Drive (Share and Collaborate)
    2.1. Open Dashboard: DQ Dashboard
    in location: SAS Content/ GTPPub /Data Management/DQ Dashboard/Dashboard
    to ensure the Dashboard data got generated correctly.

The Demo is now ready to be used.