

SAS Viya Trial

데이터 관리 가이드

데이터 엔지니어링 Task



Intro

데이터 및 AI 라이프사이클: 데이터 관리

Futurum Group의 최신 연구에 따르면, **SAS Viya**는 데이터 및 AI 팀의 **생산성을 4.6배 증가**시키는 것으로 나타났습니다.

데이터 및 AI 라이프사이클의 첫 번째 단계는 **데이터 관리**입니다. **데이터 엔지니어**는 원시 데이터(raw data)를 **평가** 및 **정제**하여 **분석 가능한 데이터 세트**로 **가공**함으로써 이후 데이터 사이언티스트가 그 다음 단계인 모델 개발로 나아갈 수 있도록 지원합니다.

본 가이드는 **SAS Viya** 환경에서 **데이터 엔지니어**가 데이터 평가 및 정제를 수행한 **과정**을 단계별로 안내합니다.

샘플 데이터 세트

샘플 데이터 세트

본 가이드에서 활용할 데이터 세트(또는 테이블)는 총 두 개입니다:

1. BANKING_ACCOUNT (10,088 rows & 19 columns)
2. BANKING_CUSTOMER (10,095 rows & 39 columns)

이 데이터 세트들의 기본 키(primary key)는 “id” 열이며, 우리의 목표 변수(target variable)는 “churn”(구독해지율)입니다.

데이터 엔지니어의 목적은 데이터를 이해, 정제 및 준비하고, 모델링에 적합한 테이블(Analytical Base Table - ABT)을 생성하는 것입니다.

생성된 테이블은 데이터 사이언티스트에게 전달되어 고객 이탈 확률이 높은 대상을 예측하도록 사용되며, 기업은 이를 바탕으로 고객 이탈 방지를 위한 조치를 취할 수 있게 됩니다.

데이터 엔지니어

데이터 정제, 준비 및 관리

개요

1. 데이터 프로파일링
2. 데이터 민감도 및
개인정보 존재 유·무 확인
3. 데이터 품질 확인
4. ETL (데이터 추출·변환·적재)
워크플로우

데이터 엔지니어 (Data Engineer)

귀하는 **데이터 엔지니어**로서 향후 활용 및 모델링에 적합한지 확인하기 위해 **기본적인 데이터 품질 검사**를 수행해야 합니다.

또한, 이 프로젝트는 민감한 정보에 접근해서는 안 되는 다양한 이해관계자들이 사용할 예정이므로, **개인정보(PII)**를 배제해 리스크를 최소화하고 규제 준수를 보장하는 것이 필수적입니다.

귀하는 **데이터 프로파일링**을 통한 품질 및 민감도 점검 및 데이터 모델링 준비를 위한 **ETL (추출·변환·적재) 워크플로우 생성** 작업을 수행할 것입니다. 생성된 워크플로우는 데이터 사이언티스트가 AI 모델을 개발하는 데 필요한 **통합 데이터 세트**를 생성합니다.

1. 데이터 프로파일링

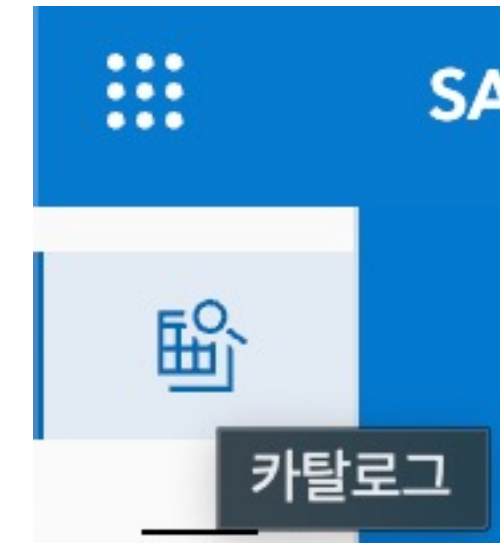
Data Profiling

SAS Information Catalog (정보 카탈로그)

우리는 데이터 품질에 대한 정보를 얻기 위해 **SAS Information Catalog (정보 카탈로그)**를 활용할 것입니다. 정보 카탈로그는 기업 전반에 분산된 정보의 메타데이터를 수집, 통합 및 보강하여 인벤토리를 생성·관리할 수 있게 해주며, 수집된 메타데이터는 필요한 데이터를 쉽고 빠르게 찾을 수 있도록 도와줍니다.

더 나아가, **정보 카탈로그**는 데이터 생성 시점, 생성자, 편집자, 최종 수정 시점 등 데이터 사용 이력을 검토할 수 있도록 지원합니다.

The screenshot displays the SAS Information Catalog interface. At the top, there's a search bar with the text "정보 에셋 검색" and a "데이터 가져오기" button. Below the search bar, there's a section titled "카탈로그 한눈에 보기" (View Catalog at a Glance) which shows several data asset categories with their respective counts: "합계 에셋" (Total Assets) 560, "데이터셋" (Datasets) 519, "리포트" (Reports) 16, "Model Studio 프로젝트" (Model Studio Projects) 9, "Lookup tables" 4, and "Rule sets" 4. To the right of this section is a "시작하기" (Get Started) area with a message: "둘러보기 SAS Information Catalog User's Guide를 살펴보세요. SAS Information Catalog의 최신 업데이트와 새로운 기능을 확인하세요!" (Browse the SAS Information Catalog User's Guide. Check the latest updates and new features of SAS Information Catalog!). Below this is a "컬렉션" (Collections) section with options for "최근" (Recent), "즐거찾기" (Favorites), and "내 데이터" (My Data). The search filter is currently set to "검색 색인: (필터 없음)" (Search Index: (No filters)).

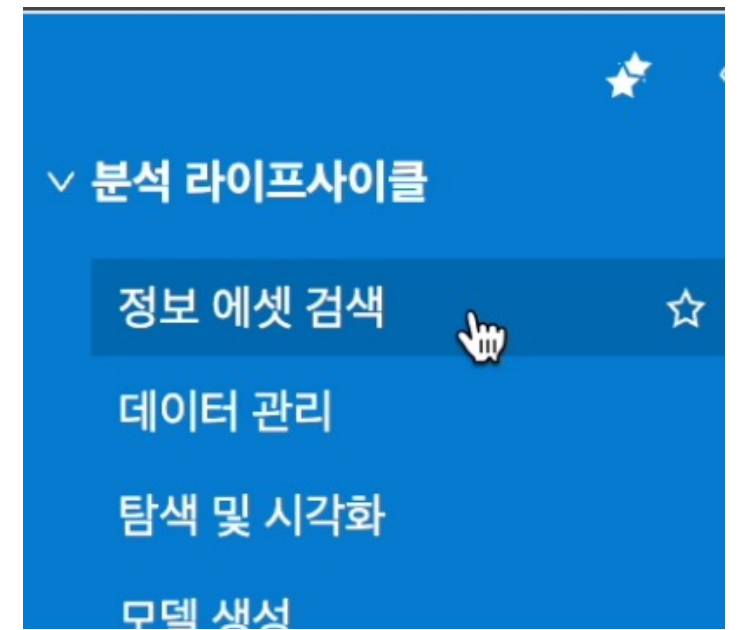


데이터 프로파일링

로딩이 완료되면 왼쪽 상단에 있는 **응용 프로그램 메뉴**¹를 클릭 후 **정보 에셋 검색**²을 클릭해주세요.



1



2

이제 검색창을 통해 저희에게 필요한 데이터 세트를 찾아봅시다. 저희가 이용할 샘플 데이터 세트의 키워드는 **“banking”**입니다.

3

A screenshot of the search results page for the keyword 'banking'. The page shows a list of data sets with columns for name, status, analysis date, and asset type. Two data sets with the name 'BANKING_ACCOUNT' are highlighted with a blue box. The right sidebar shows details for the selected data set, including its name, public status, and asset type.

이름	★	상태	분석일	에셋 유형
<input type="checkbox"/> BANKING_FOR_SCORING	☆	●	(분석되지 않음)	인메모리 데...
<input type="checkbox"/> BANKING_NEW	☆	●	(분석되지 않음)	인메모리 데...
<input type="checkbox"/> BANKING_FOR_SCORING	☆	●	(분석되지 않음)	CAS 테이블

위 사진³서 보시다시피, **중복되는 이름의 데이터 세트가 두 개 표시되는 것을 볼 수 있습니다.**

이름 옆에 번개 모양 아이콘이 있는 데이터 세트는 이미 메모리에 로드된 **‘인메모리’** 데이터 세트입니다. (또한, **‘에셋 유형’**을 통해 확인하실 수 있습니다).

저희는 이와 같은 **인메모리 데이터 세트**만 사용할 예정입니다.

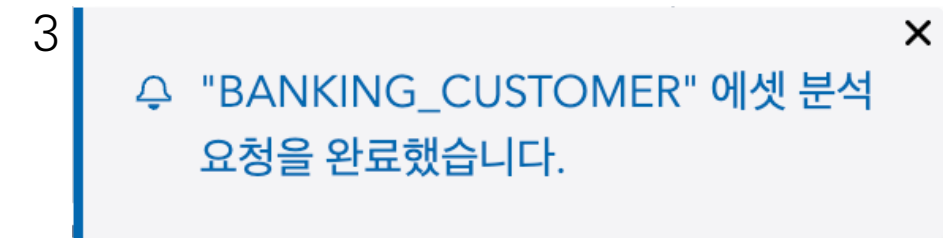
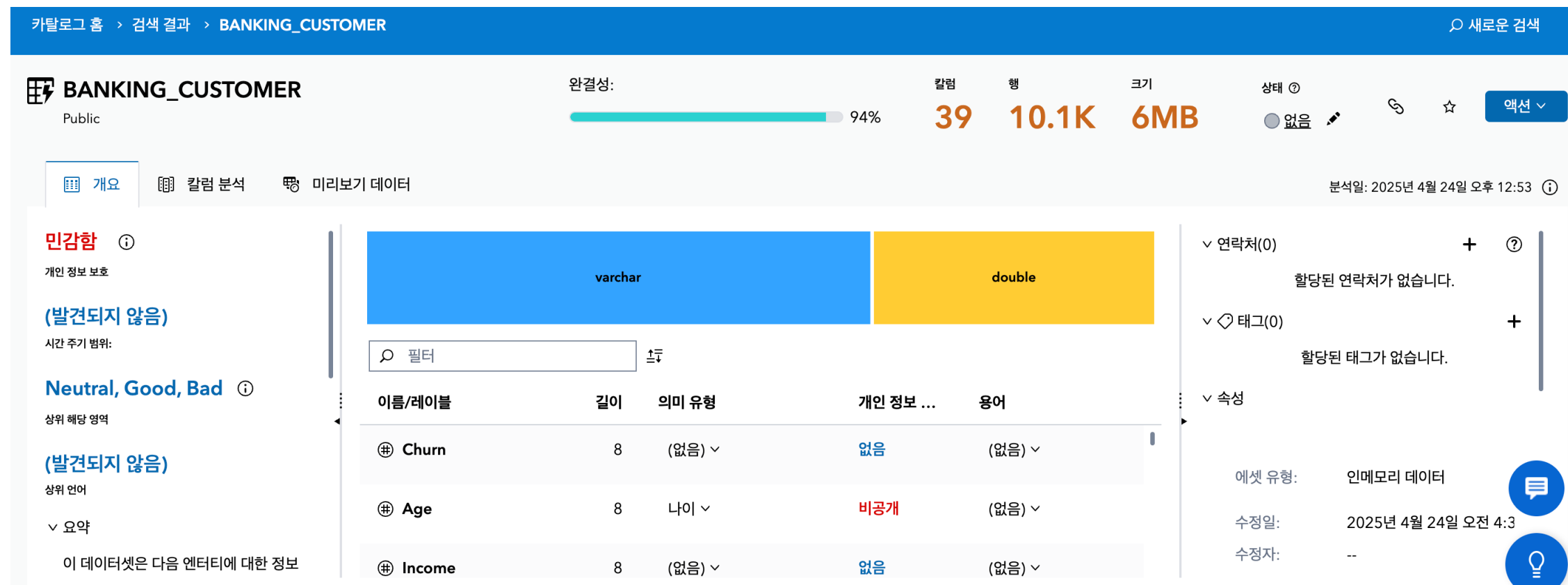
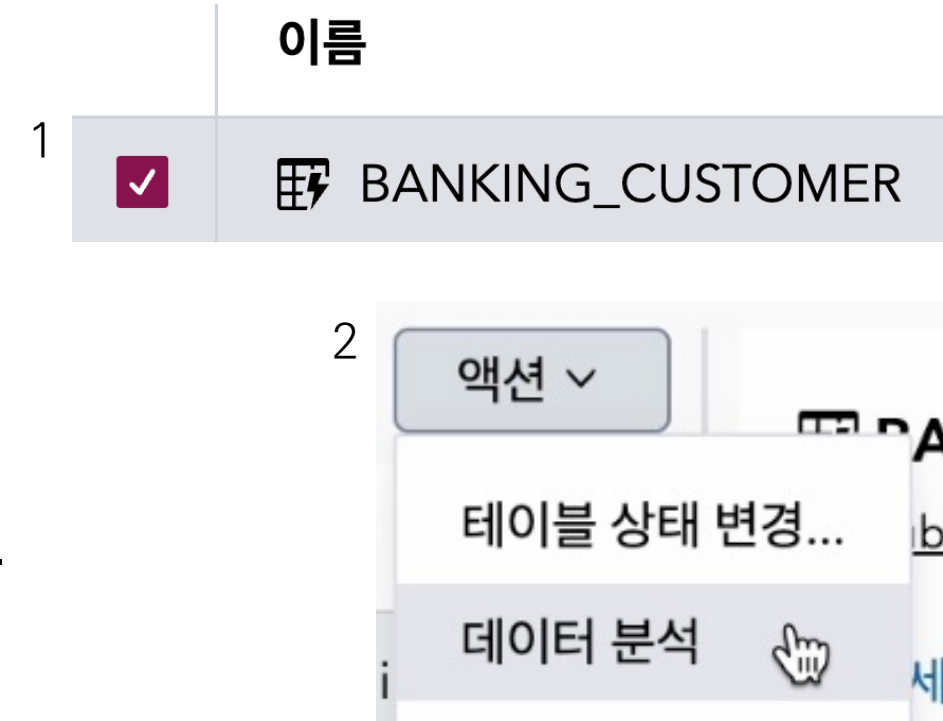
데이터 프로파일링

먼저 **BANKING_CUSTOMER** 데이터 세트에 대해 알아가봅시다.

SAS Viya에는 데이터 세트를 분석해주는 **자동화된 분석 엔진**이 내장되어 있어, 정보 카탈로그에 있는 데이터를 손쉽게 불러와 분석할 수 있습니다.

검색 결과에서 **BANKING_CUSTOMER**를 선택¹한 후, 검색 결과 우측 상단에 위치한 **액션 > 데이터 분석**²을 클릭합니다. 요청을 실행한 후 기다려줍니다.

분석이 완료³되면, 데이터 세트에 들어가 아래 사진과 같이 **분석 결과**⁴를 확인하실 수 있습니다. 이제부터는 주요 지표 및 통계를 통해 결과를 더 자세히 살펴보겠습니다.



분석 완료 시, 위와 같은 알림이 화면에 게시됩니다.

4

2. 데이터 민감도 및 개인정보 존재 유·무 확인 Data Sensitivity & PII Checks

칼럼별 데이터 민감도 알아보기

먼저 화면 중앙¹에 위치한 “개인 정보 보호”를 통해 칼럼별 데이터 민감도를 확인하실 수 있습니다.
화면 왼쪽²에서는 중요 데이터 및 이상값이 탐지된 결과를 자연어로 요약한 내용을 확인하실 수 있으며, 이 두 설명 모두 SAS Viya에 의해 자동 생성됩니다.

“개인 정보 보호”는 총 네 가지 카테고리로 분류되며, 이런 결과를 바탕으로 데이터 엔지니어는 고객의 개인·민감 데이터를 보호하고, 필요 시 이상점(outlier)을 조사 및 처리하는 조치를 취해야 합니다.

개인 정보 보호 - 4가지의 카테고리 (오름차순):

1. **없음**: 칼럼에 개인정보 없음
2. **후보**: 칼럼에 개인정보가 존재할 가능성 (위치 정보 등)
3. **비공개**: 칼럼에 개인정보 포함 (이름, 성, 나이 등)
4. **민감함**: 칼럼에 민감정보 포함 (혼인 여부 등)

The screenshot shows the SAS Viya interface for the 'BANKING_CUSTOMER' table. The table has 39 columns, 10.1K rows, and 6MB size. The completion status is 94%. The interface displays a table of columns with their respective data types and sensitivity levels. A blue box highlights the '개인 정보 보호' (Data Sensitivity) column, and a red box highlights the '민감함' (Sensitive) status for the 'Gender' and 'Marital_status' columns. A blue box also highlights the summary text on the left side of the interface.

이름/레이블	길이	의미 유형	개인 정보 보호 ↓	용어
Gender	6	성별 ↓	민감함	(없음) ↓
Marital_status	9	혼인 여부 ↓	민감함	(없음) ↓
Age	8	나이 ↓	비공개	(없음) ↓
Education	11	성 ↓	비공개	(없음) ↓
Savings_behavior	9	성 ↓	비공개	(없음) ↓

데이터 민감도 확인 후 데이터 세트 상태 변경하기

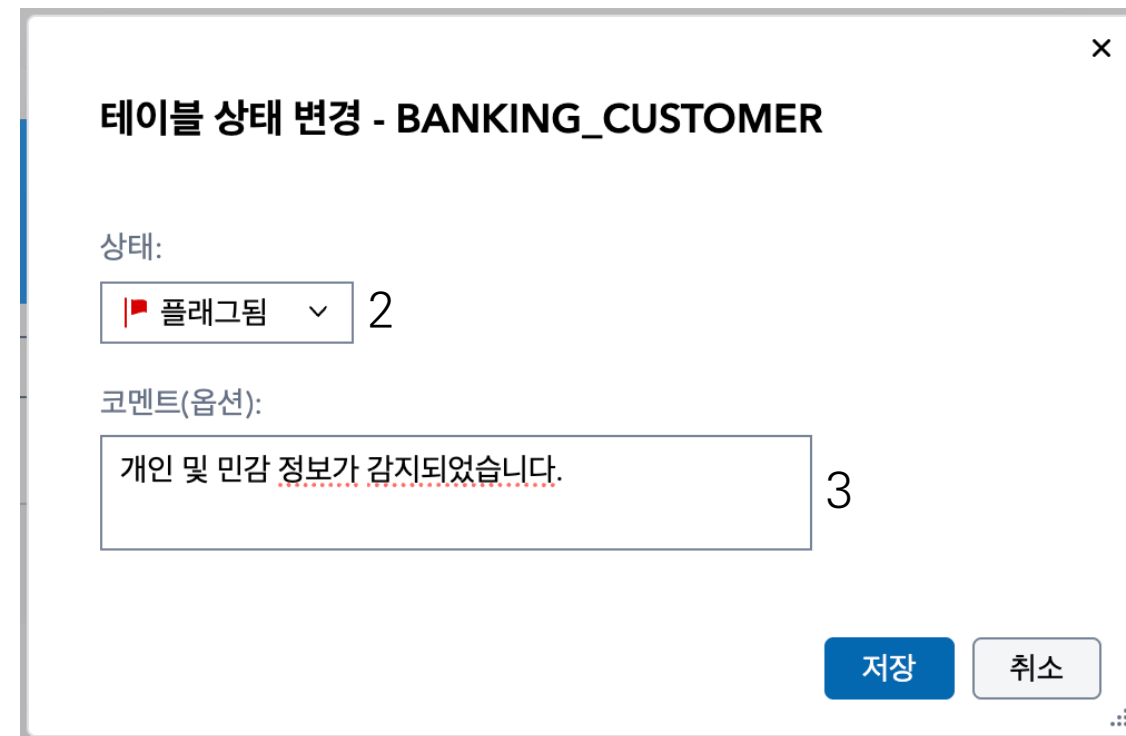
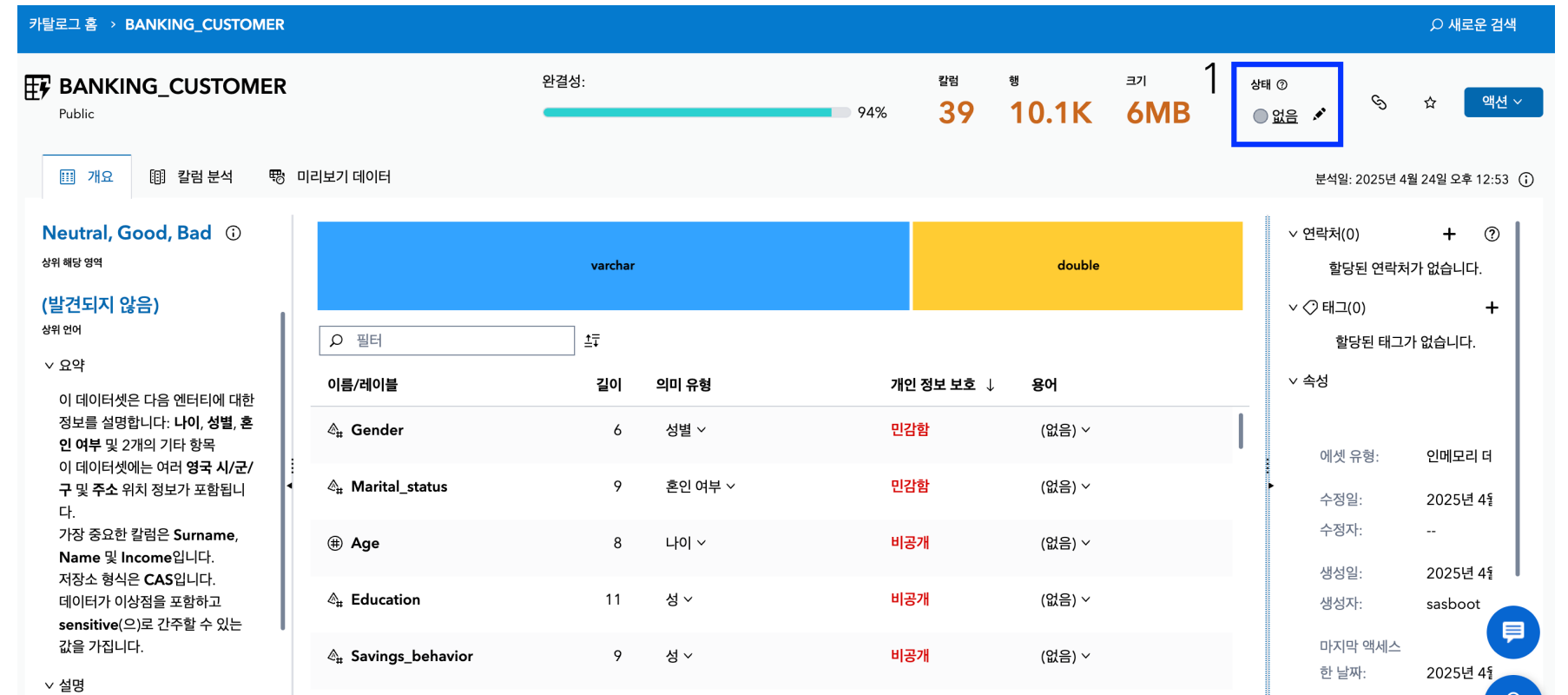
저희는 데이터 세트에 개인 및 민감 정보가 포함된 것을 확인했습니다.

이제 저희는 접근 권한이 없는 타 부서와의 공유를 방지하기 위해 확인 결과를 데이터 엔지니어 동료들에게 안내해야 합니다.

“상태” 탭에서 연필 아이콘¹을 클릭하여 경고 플래그를 설정합니다.

설정 화면에서 상태²를 “플래그됨”으로 선택하고 “개인 및 민감 정보가 감지되었습니다”와 같은 코멘트³를 추가하여 플래그 설정 이유를 기재합니다.

설정이 완료되면 이 플래그는 SAS Viya에서 해당 데이터 세트에 접근할 수 있는 모든 사용자에게 표시됩니다.



3. 데이터 품질 확인

Data Quality Check

데이터 품질 확인하기 (칼럼 분석 > 데이터 품질 측도)

지금까지 데이터의 민감도를 확인했다면, 다음 단계는 각 칼럼별 데이터의 품질을 알아보는 것입니다.

데이터의 품질을 확인하기 위해 칼럼 분석¹ > 데이터 품질 측도²를 클릭합니다.

1: 칼럼 분석 탭 클릭

2: 데이터 품질 측도 탭 클릭

3: ID 컬럼의 고유성 99% 확인

#	이름	완결성	고유성	가장 흔한 값	가장 드문 값
32	Product_...	100%	0%	/	o
33	Survey_r...	100%	0%	No	Yes
34	Custome...	75%	0%	Good	Bad
35	Cross_se...	100%	0%	No	Yes
36	Market_c...	100%	0%	Neutral	Bad
37	Name	100%	1%	Liam	Dylan
38	Surname	100%	1%	Smith	Harrison(1개 이상)
39	Id	100%	99%	9856	1

위 예시³에서 “ID”가 기본 키임에도 고유성이 100%가 아닌 것으로 보아, 데이터 세트에 중복 레코드가 존재한다는 것을 알 수 있습니다.

클릭하시면 아래와 같은 기본적인 데이터 품질 관련 사항들을 확인하실 수 있습니다:

완결성: 데이터 값이 누락 없이 완전히 채워져 있는 정도

고유성: 각 레코드나 값이 중복 없이 고유하게 나타나는 정도

가장 흔한 값·가장 드문 값

패턴 수: 칼럼에서 발생하는 고유 단어 또는 문자 패턴의 수

의미 유형: 칼럼에 포함될 가능성이 있는 데이터의 분류를 식별

개인 정보 보호: 칼럼에 개인과 연관될 수 있는 잠재적인 개인 정보가 포함되어 있는지를 나타냄

데이터 품질 확인하기

더 세부적인 사항을 보기 위해
대상 칼럼을 클릭합니다.

# ↑	이름
3	# Income

Income
레이블: (없음)

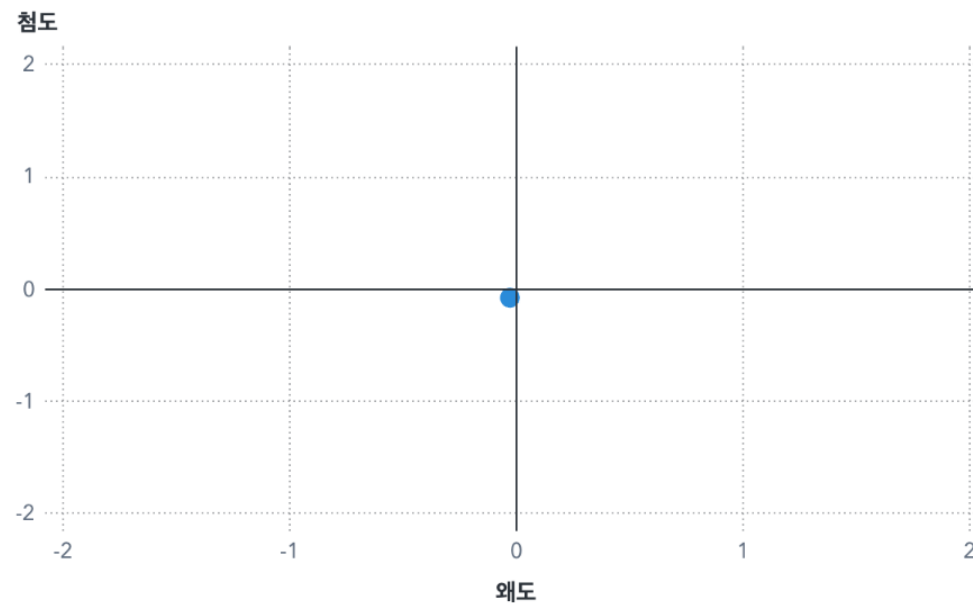
의미 유형 (없음) ↓ 개인 정보 보호 **없음** 주기 후보 **아니요** ⓘ

데이터 품질



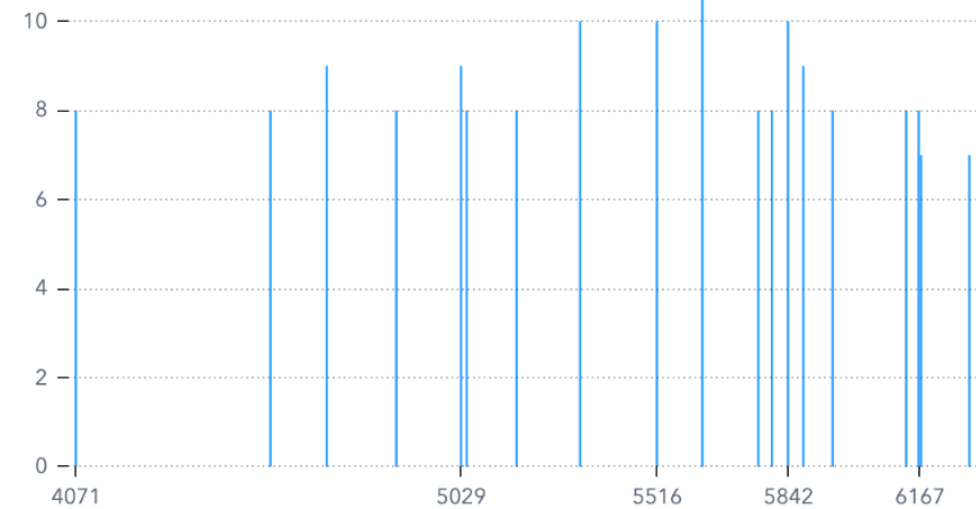
정규성 편차

왜도 **-0.02495** 첨도 **-0.07405** 표준편차 **1308.202**

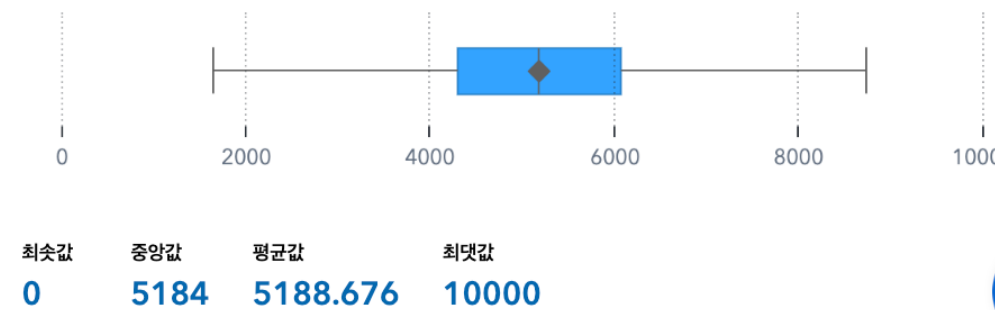


빈도 분포

이상점 숨기기 **최고** 최저



전체 분포



클릭하시면 아래와 같은 데이터 품질에 관한
세부 사항들을 상세히 확인하실 수 있습니다:

왜도: 데이터 분포의 좌우 비대칭도를 표현하는 척도

첨도: 분포가 정규분포보다 얼마나 뾰족하거나 완만한지의 정도를 나타내는 척도

완결성: 데이터 값이 누락 없이 완전히 채워져 있는 정도

고유성: 각 레코드나 값이 중복 없이 고유하게 나타나는 정도

고유타값: 특정 칼럼에서 중복을 제외한 고유 값의 총 개수

숫자형 변수의 경우 통계 지표가,
문자형 변수의 경우 패턴 및 빈도가 표시됩니다.

데이터 품질 확인하기 (칼럼 분석 > 기술 측도)

이외에도 각 칼럼별 데이터 품질에 대한 인사이트를 더 얻고자 칼럼 분석 탭에 있는 “기술 측도”를 활용합니다.

예시로, 칼럼 분석 > 기술 측도 클릭 후 결측값(missing value)을 기준으로 정렬합니다.

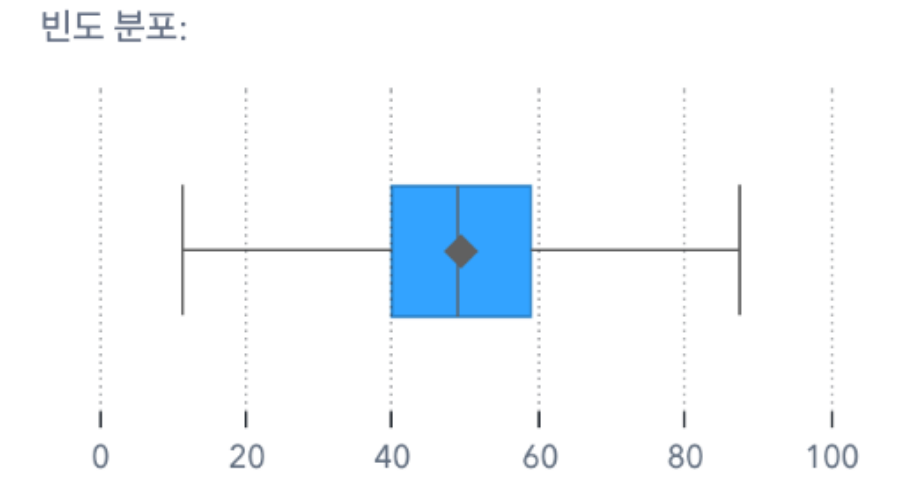
옆에 게시된 칼럼들¹의 경우 값이 누락된 케이스가 존재한다는 것을 알 수 있습니다.

원하는 데이터 칼럼을 선택하시면 화면 우측에 빈도, 분포도, 분위수 등을 확인하실 수 있으며,

우측 상단의 아이콘³을 클릭하시면 선택한 칼럼의 이상점(outlier) 개수를 확인하실 수 있습니다.

1 이름	결측값 ↓
△ Loyalty_program	5,615
△ Life_event_marriage	4,972
△ Socialmedia_usage	3,287
△ Financial_literacy	2,463
△ Customer_sentiment	2,446
△ Digital_usage	1,127
⊕ Age ⋯	192

⊕ Age ⋯ 2



√ 분위수

최솟값:	0
25%:	40
50%:	49
75%:	59
최댓값:	100

⋯ Age은(는) 분석에 영향을 줄 수 있는 81개의 이상점을 포함합니다. ⋯ 3

책 ↑	이상점 ↓	빈도 분포:
--		
--		
--		
--		

√ 분위수

유형:	double
출력형식:	--
길이:	8
주 키 후보: ②	아니요
논리적 유형: ②	Interval
의미 유형: ②	나이 ∨
개인 정보 보호: ②	비공개

4. ETL(데이터 추출·변환·적재)

워크플로우

ETL Flow

워크플로우 생성하기

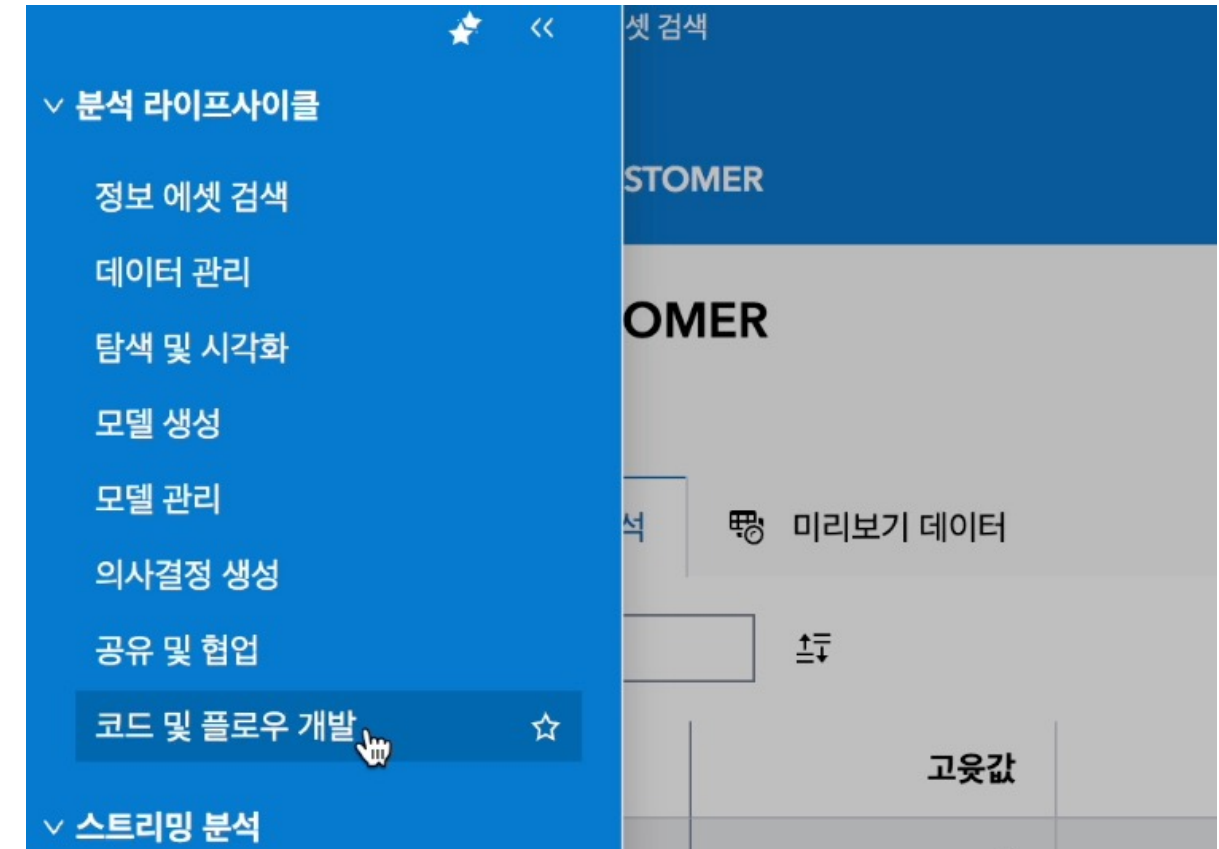
지금까지 데이터 세트의 품질을 점검했다면, 이제 로우코드/노코드 환경에서 손쉽게 구현할 수 있는 **커스텀 ETL (데이터 추출·변환·적재) 워크플로우**를 구축하여 기존의 **반복적인 프로세스를 자동화**해 보겠습니다.

ETL 워크플로우의 최종 목표는

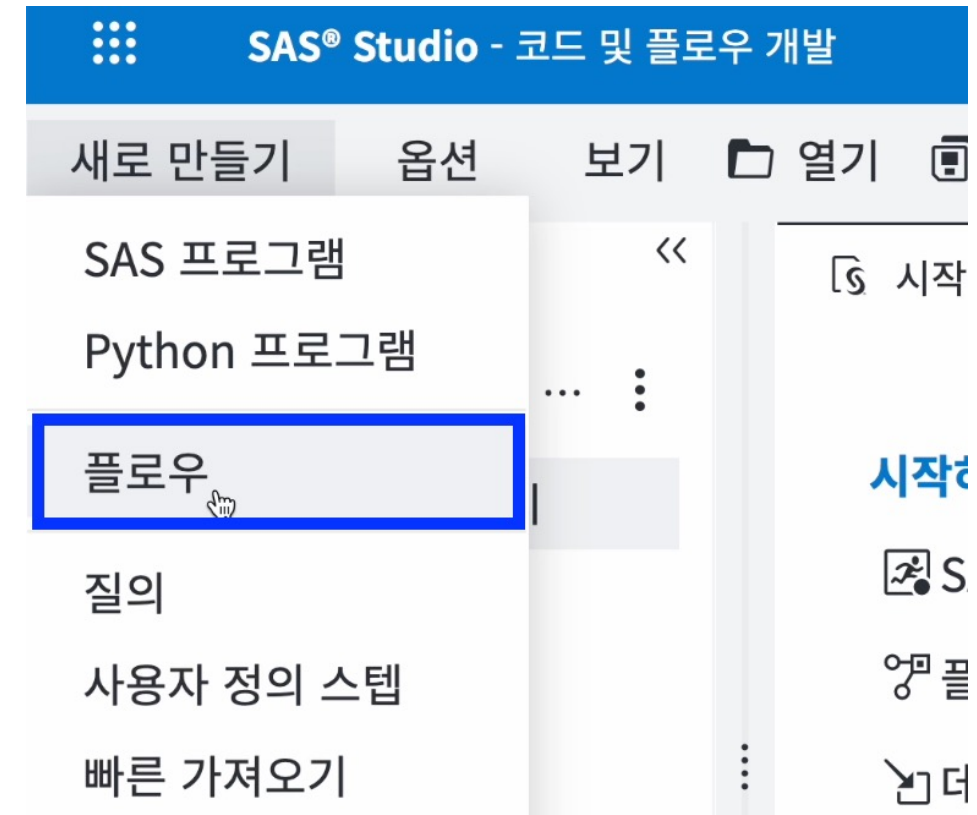
데이터 사이언티스트가 AI 모델을 개발하기 위해 필요한 **통합 데이터 세트(ABT 테이블)**를 생성하는 것입니다.

1. 왼쪽 상단에 있는 **응용 프로그램 메뉴**를 클릭한 후 **코드 및 플로우 개발¹**을 선택해주세요.
2. **새로 만들기 > 플로우²**를 클릭해 새로운 플로우 파일을 생성하세요.

1



2



ETL 워크플로우

저희는 워크플로우를 통해 저희에게 주어진 두 데이터 세트 (BANKING_ACCOUNT & BANKING_CUSTOMER)를 데이터 세트의 기본 키인 ID를 이용하여 합칠 예정입니다.

하지만, 방금 전 데이터 세트의 품질에 대해 알아가던 도중 저희가 활용할 데이터 세트에 **중복 레코드**가 존재한다는 것을 확인했습니다¹.

더 나아가, 데이터 세트에는 **민감한 정보가 포함되어 있어**², 배포 전에 특정 칼럼 값을 마스킹(Masking)하는 과정을 거쳐야 합니다.

따라서, 저희는

1. 두 데이터 세트를 불러온 후
2. 중복되는 레코드를 제거하고
3. ID를 이용하여 두 데이터 세트를 (JOIN) 합친 후에
4. 개인정보 및 민감정보 마스킹 (Data Masking)을 거쳐
5. 변수 구관화 (Binning) 후
6. 최종적인 ABT 테이블을 생성할 것입니다.

1

이름	완결성	고유성
⊕ Product_...	100%	0%
△ Survey_r...	100%	0%
△ Custome...	75%	0%
△ Cross_se...	100%	0%
△ Market_c...	100%	0%
△ Name	100%	1%
△ Surname	100%	1%
⊕ Id	100%	99%

2

이름/레이블	길이	의미 유형	개인 정보 보호 ↓
△ Gender	6	성별 ∨	민감함
△ Marital_status	9	혼인 여부 ∨	민감함
⊕ Age	8	나이 ∨	비공개
△ Education	11	성 ∨	비공개
△ Savings_behavior	9	성 ∨	비공개

워크플로우 내 스윘레인 생성 및 활성화

플로우는 스윘레인을 통해 프로세스를 시각적으로 분할하며, 각 스윘레인 안에 실행 가능한 스니펫(snippet) 또는 노드(node)를 배치합니다.

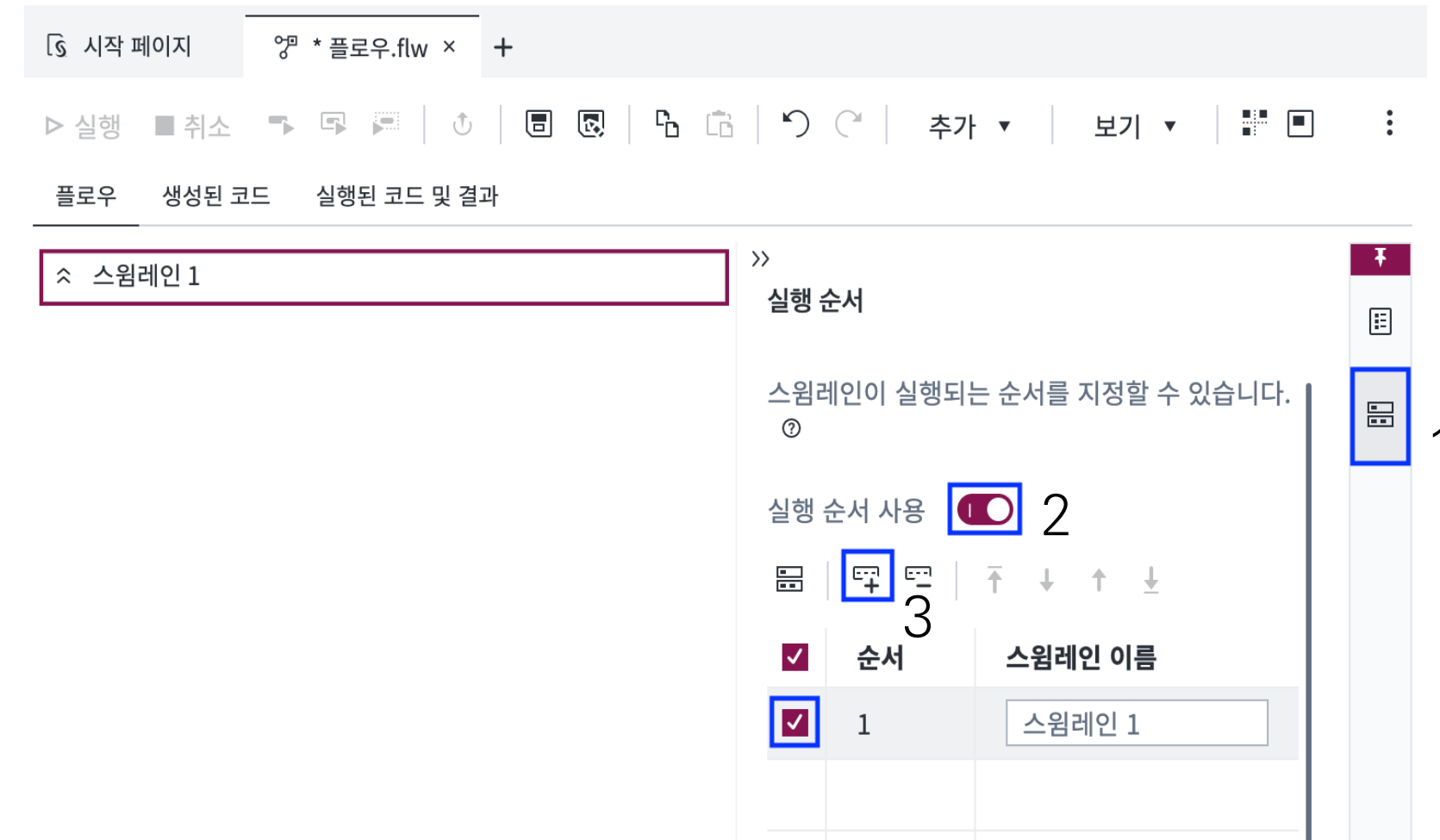
새로운 플로우 파일을 생성한 뒤, 우측 중앙의 **실행 순서**¹를 클릭하고 팝업에서 **실행 순서 사용**²을 활성화하세요.

활성화^a되면 스윘레인 1이 자동 생성되며, 추가 스윘레인이 필요할 경우 **플러스 아이콘**³을 눌러 원하는 만큼 추가할 수 있습니다.

이제부터 활성화된 스윘레인 1에 스니펫 및 노드를 추가합니다.

스니펫이란 워크플로우 내에서 실행되는 **개별 코드 블록**을 말하며, 스니펫을 이용하여 SAS 코드를 프로그램에 빠르게 삽입할 수 있으며, 필요에 맞게 손쉽게 맞춤화할 수 있습니다.

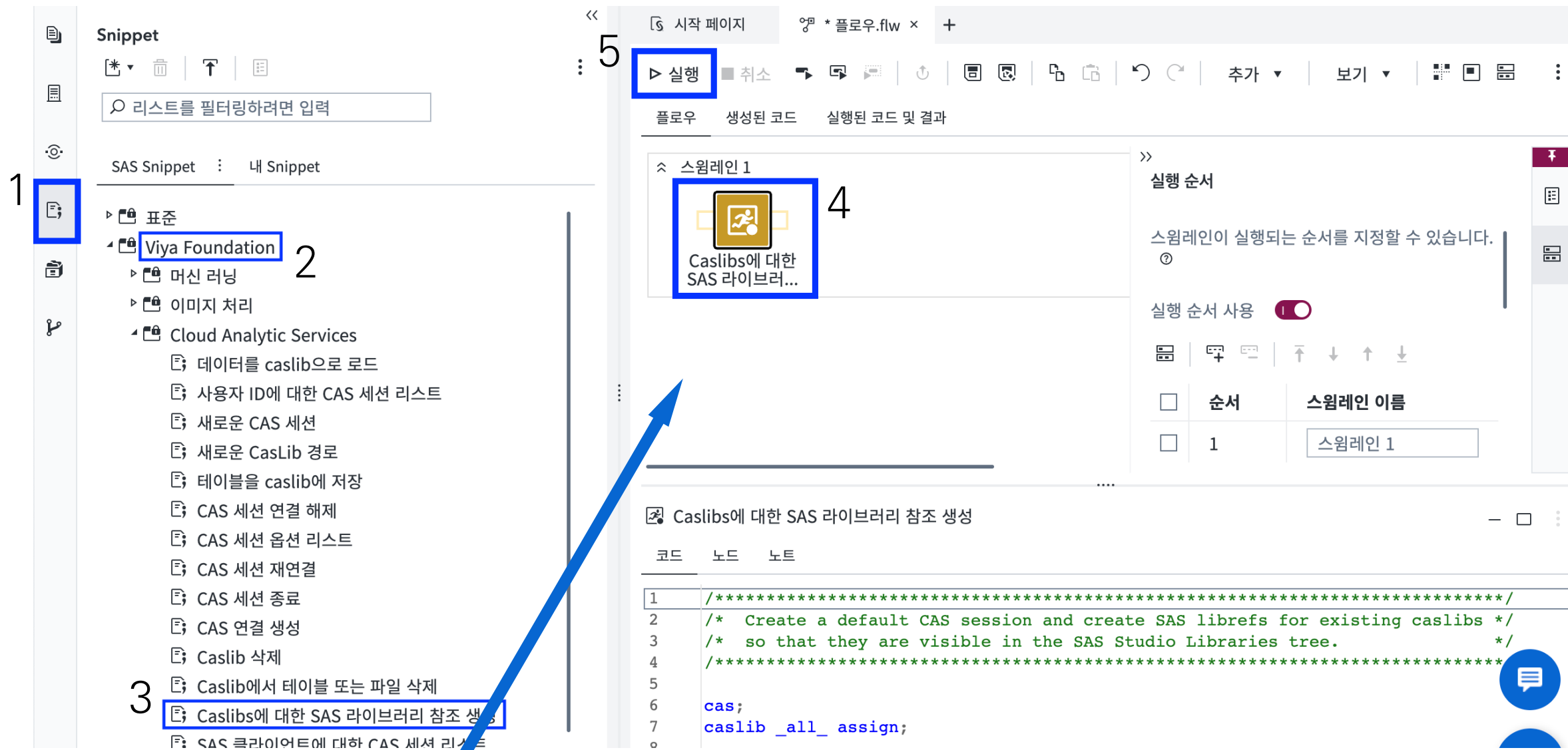
노드란 워크플로우에서 하나의 **구체적인 작업 단위**를 의미하며, 데이터 로드·변환·병합·마스킹 등 특정 기능을 수행하는 요소입니다.



CAS와 연결

스웸레인 1이 플로우에 추가된 시점에서 가장 첫 번째 단계는 SAS Viya의 인메모리 엔진인 Cloud Analytic Services(CAS)에 연결하여 새로운 세션을 시작하는 것입니다.

Snippet¹ > Viya Foundation² > Cloud Analytic Services³에서 Caslibs에 대한 SAS 라이브러리 참조 생성 스니펫을 찾아 스웸레인 1에 드래그 앤 드롭하세요.



스니펫 아이콘⁴을 클릭하시면 자동으로 생성된 코드를 확인할 수 있습니다.

실행⁵을 눌러 플로우를 실행하세요.



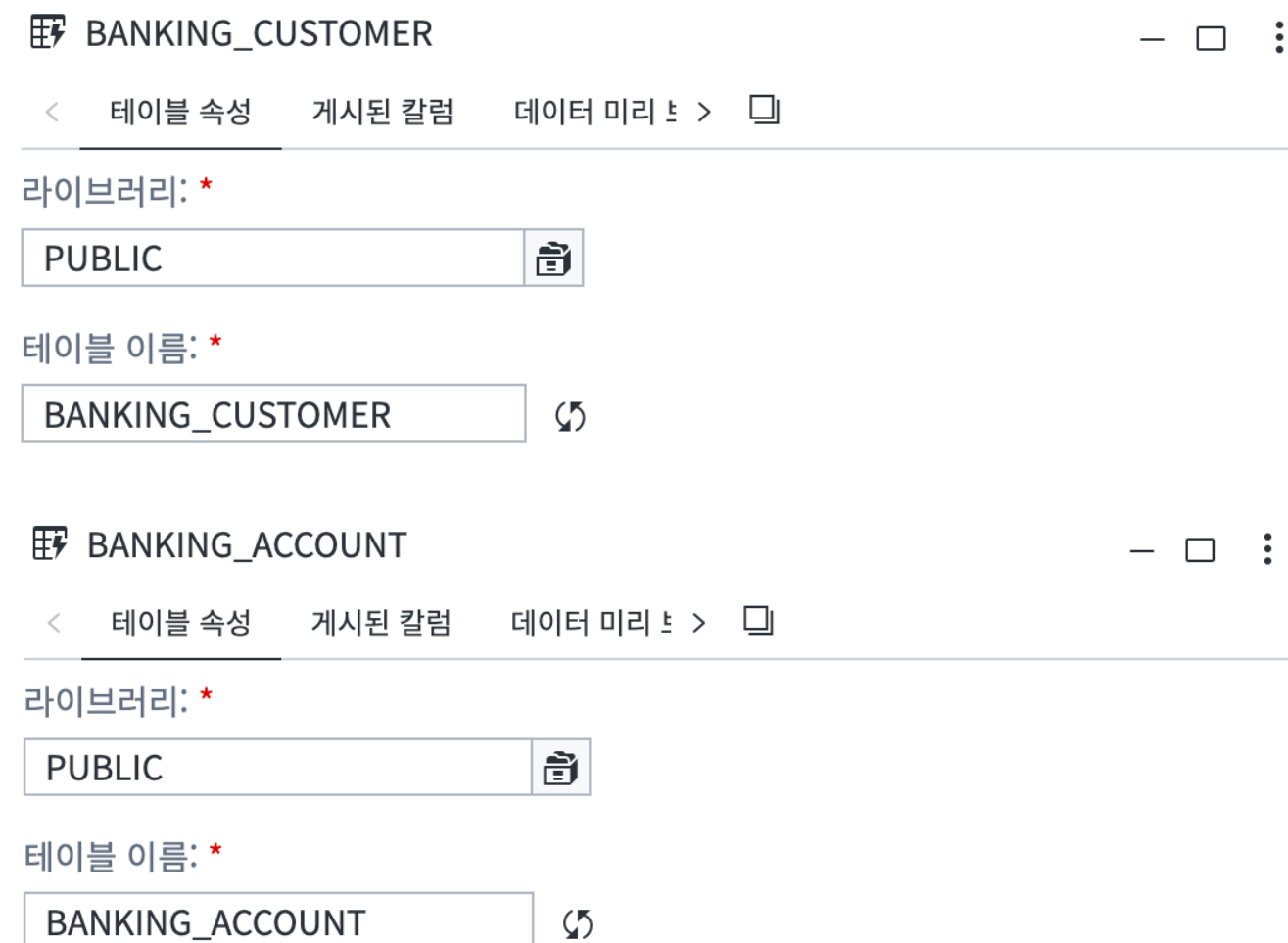
스니펫 아이콘에 체크 마크⁶가 있을 시 해당 스니펫이 정상적으로 실행되었다는 것을 의미합니다.

1. 데이터 세트 불러오기

CAS 엔진을 활용하여 데이터 세트를 불러옵니다.
새로운 스웬라인 (스웬라인 2) 추가 후

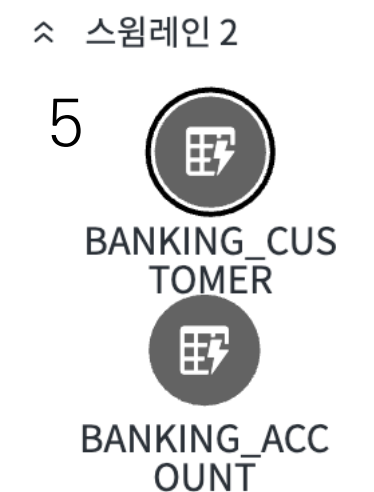
노드¹ > 데이터 (입력 및 출력)²에서
테이블³ 노드를 찾아 스웬라인 2에 드래그 앤 드롭하세요.

두 개의 데이터 세트를 불러와야 하기 때문에
총 두 개의 테이블 노드를 추가하세요.



각 테이블 노드를 클릭하여 불러올 데이터 세트의
라이브러리 및 이름을 위와 같이 기재해주세요.

설정 후 테이블 노드에 번개 아이콘⁵이 추가되며,
이는 데이터 세트를 불러올 준비가 완료됨을
의미합니다.



2. 중복 데이터 제거

두 데이터 세트가 준비되었습니다.

이제 중복 데이터를 제거해 보겠습니다.

1. **데이터 변환¹** > **중복 제거²** 노드를 찾아 스웬라인 2로 드래그하세요.
 - 데이터 세트마다 하나씩, 총 **두 개의 중복 제거** 노드를 추가합니다.
2. 각 **데이터 세트 아이콘** 테두리를 클릭한 채로 **중복 제거** 노드로 **드래그해 연결**하세요.
3. 두 데이터 세트 연결 후 **중복 제거**를 클릭하여
 - **중복 제거³**란에 “모든 칼럼에서 중복 제거”가 되어있는지
 - **출력란⁴**에 “기존 출력 테이블 바꾸기”가 되어있는지
 확인 후 **스웬라인 2**를 우클릭 > **스웬라인 실행⁵**을 선택해 중복 제거를 실행하세요.

The screenshot illustrates the SAS Studio workflow for removing duplicate data. It is divided into two main panels: the left 'Steps' panel and the right 'Flow' panel.

Steps Panel (Left):

- Shows a search bar: "리스트를 필터링하려면..."
- Under "SAS 스텝 : 공유", the "데이터 변환" (Data Transformation) category is expanded, showing options like "데이터 순위화", "데이터 전치", "마스킹 데이터", "정렬", and "중복 제거" (Duplicate Removal).
- The "중복 제거" node is highlighted with a blue box and labeled "2".

Flow Panel (Right):

- Shows a workflow with two swimlanes: "스웬라인 1" and "스웬라인 2".
- In "스웬라인 2", two data source nodes are connected to "중복 제거" nodes: "BANKING_CUS TOMER" to "중복 제거" and "BANKING_ACC OUNT" to "중복 제거 2".

Configuration Dialog (Bottom):

- Shows the configuration for the "중복 제거" node.
- Option 3: "중복 제거" is selected, with "모든 칼럼에서 중복 제거" checked.
- Option 4: "출력" (Output) is selected, with "기존 출력 테이블 바꾸기" checked.

Execution Panel (Bottom Right):

- Shows the "스웬라인 2" context menu with "스웬라인 실행" (Execute Swimlane) highlighted with a blue box and labeled "5".

3. 데이터 병합 (JOIN)

중복 데이터 제거가 완료되면, 다음 단계는
두 데이터 세트를 기본 키인 "ID"를 이용하여 합치는 겁니다.

1. 데이터 변환¹ > 질의² 노드를 찾아 스윙레인 2로 드래그하세요.
2. 중복 제거의 우측 앵커(네모)³를 클릭한 채로 질의⁴ 노드로 드래그해 연결합니다.
3. 두 중복 제거 노드 연결 후 질의를 클릭하여
 1. 선택⁵ > 칼럼⁶에서 왼쪽 패널의 두 데이터 세트⁷ (t1, t2)를 선택 창으로 드래그하세요.
 2. 조인⁸에서 t1과 t2 사이의 조인 아이콘⁹ 클릭 후 레프트 조인¹⁰을 선택합니다.
 - LEFT JOIN은 t1(BANKING_CUSTOMER)의 모든 레코드를 유지한 채 t2(BANKING_ACCOUNT)와 병합합니다.
4. 설정을 완료한 후 질의 노드 우클릭 > 노드 실행¹¹을 선택해 데이터 병합을 실행합니다.



1. 데이터 변환

- 데이터 순위화
- 데이터 전치
- 마스킹 데이터
- 정렬
- 중복 제거

2. 질의

스윙레인 2

BANKING_CUS TOMER → 중복 제거

BANKING_ACC OUNT → 중복 제거 2

중복 제거 → 질의

중복 제거 2 → 질의

질의

질의

옵션 노드 노트

칼럼

5. 선택

6. 칼럼

7. t1 (중복 제거), t2 (중복 제거 2)

8. 조인

9. t2(중복 제거 2)

10. 레프트 조인

11. 노드 실행

워크플로우 및 중간 데이터 저장하기

지금까지의 워크플로우 과정을 저장하기 위해 **다른 이름으로 저장**¹을 클릭하고 저장 경로로 SAS 콘텐츠 > Users > My Folder 로 선택한 뒤 저장해줍니다.

다음 단계는 **중간 데이터 세트를 개인 라이브러리에 저장하는 것입니다.** **노드 > 데이터 (입력 및 출력)**에서 **테이블** 노드를 찾아 **스웸레인 2**에 드래그 앤 드롭하세요.

현재 가공된 데이터 세트는 **최종 ABT 테이블이 아닌 체크포인트의 역할을 하기 때문에** 개인 유저만 볼 수 있는 **CASUSER** 라이브러리에 저장할 것입니다.

따라서, **테이블 속성**²에서 라이브러리를 **CASUSER**³로 설정합니다. 테이블 이름은 “Banking_Transformed”로 설정하겠습니다.

설정을 완료한 후 **테이블** 노드 우클릭 > **종료 노드 실행**⁴을 클릭하여 **중간 데이터 세트를 CASUSER** 라이브러리에 저장해줍니다.

보시다시피, “Banking_Transformed”가 CASUSER에 (**라이브러리**⁵ > **연결된 라이브러리**⁶ > **CASUSER**⁷) 저장된 것을 확인하실 수 있습니다.

1 다른 이름으로 저장

SAS 스텝 : 공유

스웸레인 2

BANKING_CUS TOMER 중복 제거

BANKING_ACC OUNT 중복 제거 2

질의

마스킹 데이터

테이블

2 Banking_Transformed

테이블 속성 옵션 게시된 칼럼 데이터 미리 보기 노드 노트

3 라이브러리: *

CASUSER

테이블 이름: *

Banking_Transformed

4 종료 노드 실행

5

라이브러리

연결된 라이브러리 6

7 CASUSER

BANKING_TRANSFORMED

FORMATS

MAPS

MAPSGFK

MAPSSAS

5. 변수 구간화 (Binning)

마지막 단계는 민감 변수인 "Age"를 구간화(binning)하여 개인정보 노출을 방지하되 데이터의 분석적 유용성은 유지하는 겁니다.

변수 구간화는 연속형 변수를 특정 구간으로 나누어 범주형 또는 순위형 변수로 변환하는 방법을 말합니다. 저희는 변수 Age를 총 다섯 구간으로 나눌 겁니다.

1. 데이터 준비¹ > 구간화²를 찾아 스왑레인 2로 드래그하세요.
2. Banking_Transformation 데이터 세트 아이콘 테두리를 클릭한 채로 구간화 노드로 드래그해 연결하세요.
3. 이제 구간화 노드를 클릭하여 구간화 설정을 완료합니다.

1. 구간화할 칼럼 선택하기

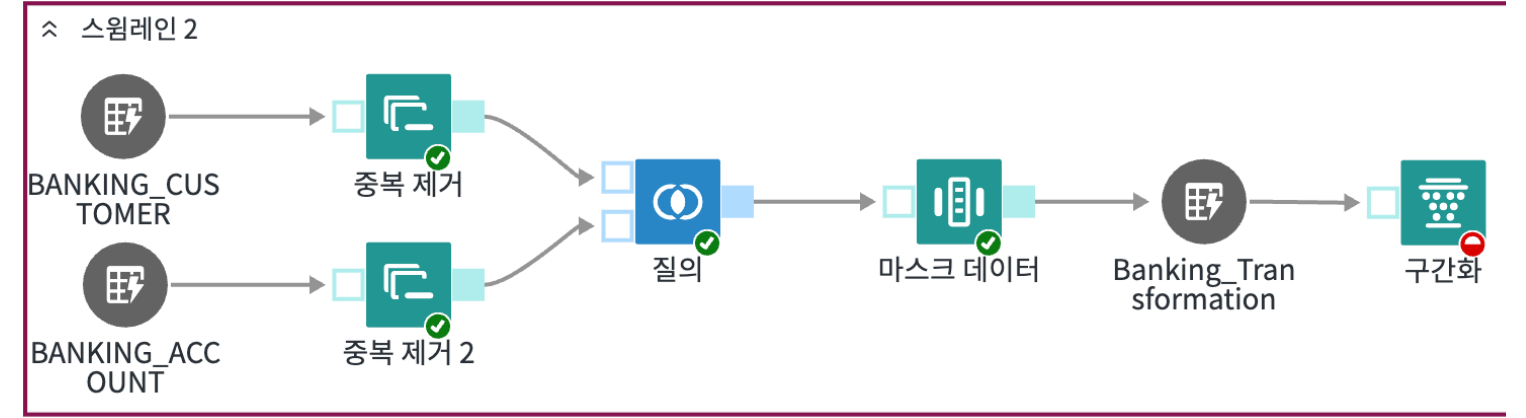
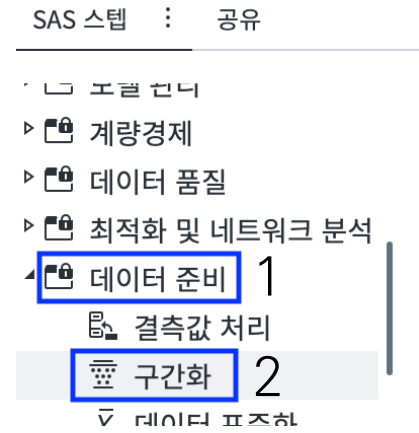
- 데이터³ > 구간화할 Interval 입력값⁴에 플러스 사인을 누른 후 리스트에서 Age를 선택해줍니다.

2. 구간 개수 설정하기

- 옵션⁵ > 구간 수⁶ 값을 5로 설정해줍니다.

3. 출력 설정하기

- 출력⁷ > 구간화된 데이터 저장⁸ 및 기존 출력 테이블 바꾸기⁹를 활성화하고 선택한 변수¹⁰를 이용하여 모든 변수(칼럼)을 가져올 수 있도록 합니다.



3 데이터 옵션 출력 노트 노트

입력 데이터 필터링:

구간화할 Interval 입력값: * ↑ ↓ 🗑️ + 4

Age

구간화 5

데이터 옵션 출력

구간 수: * 6 5

방법: * 버킷 구간화(기본)

구간화 7

데이터 옵션 출력 노트 노트

구간화된 데이터 저장 8

노트: 플로우에서 이 스텝을 실행할 때 스텝 노트에서 이 옵션 테이블에 대한 출력 포트를 추가해야 합니다.

기존 출력 테이블 바꾸기 9

입력 CAS 테이블의 변수 포함:

모든 변수

분석에 사용된 변수

변수 없음

선택한 변수 10

플러스 아이콘을 눌러 출력 값에 포함할 변수들을 추가합니다.

다음 변수 포함: * ↑ ↓ 🗑️ +

Recency

Regularity

6. ABT 테이블 생성

변수 구간화 설정까지 완료했다면, 이제 남은 단계는 **최종 ABT 테이블을 생성 및 저장하는 것입니다.**

최종 ABT 테이블은 데이터 사이언티스트가 분석을 위해 활용할 수 있도록 개인 라이브러리가 아닌 PUBLIC 라이브러리에 저장할 것입니다.

1. 구간화 노드 우클릭 후 **출력 포트 추가**¹를 선택합니다 (우측 앵커(네모)가 추가된 것을 확인하실 수 있습니다²).
2. **노드 > 데이터 (입력 및 출력)**에서 **테이블 노드**를 찾아 **스왈레인 2**에 드래그 앤 드롭하세요.
3. **구간화 노드의 우측 앵커(네모)**를 클릭한 채로 **테이블**로 드래그해 연결합니다.
4. **테이블 노드를 클릭** 후 아래와 같이 설정해줍니다:

BANKING_ABТ

테이블 속성 옵션 게시된 칼럼 데이터 미리 보기 노드 노트

라이브러리: *

PUBLIC

테이블 이름: *

BANKING_ABТ

노드 실행
시작 노드 실행
종료 노드 실행

마지막으로 실행된 코드로 이동
마지막으로 실행된 로그로 이동

포트 펼치기
1 출력 포트 추가
잘라내기

2

구간화

5. 설정을 완료한 후 **테이블 노드 우클릭 > 종료 노드 실행**³을 클릭하여 중간 데이터 세트를 CASUSER 라이브러리에 저장해줍니다.

구간화

BANKIN

3

노드 실행
시작 노드 실행
종료 노드 실행

PUBLIC

- ADMISSIONS_1_Q1
- AUSTIN_DAY_OF_WEEK_STA...
- AUSTIN_INTAKE_AND_OUTC...
- AUSTIN_MERGED
- BANK_CUSTOMERS
- BANK_CUSTOMERS_1
- BANK_CUSTOMERS_CLEANS...
- BANKING_ABТ**

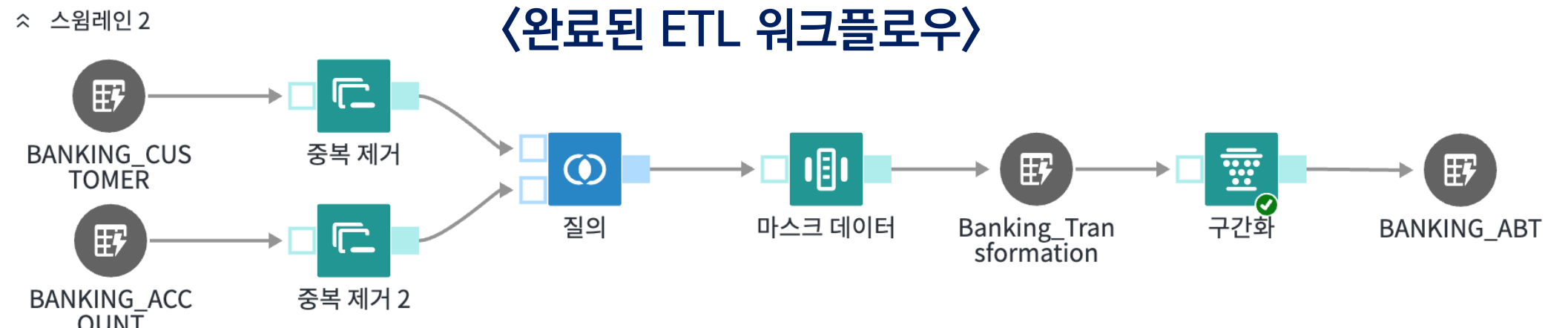
보시다시피, **BANKING_ABТ**가 **PUBLIC**에 저장된 것을 확인하실 수 있습니다.

ABT 테이블 생성을 끝으로 데이터 엔지니어로서의 작업이 완료되었습니다.

BANKING_ABТ 테이블 행: 10000 킬

표현식 입력

	⊕ Recen...	⊕ Regularity	⊕ Monetary	⊕ Amount_avg	⊕ Churn
1	5	3	5	50	1
2	6	4	5	53	1
3	4	4	5	43	0
4	6	4	5	39	0



End of the Guide

