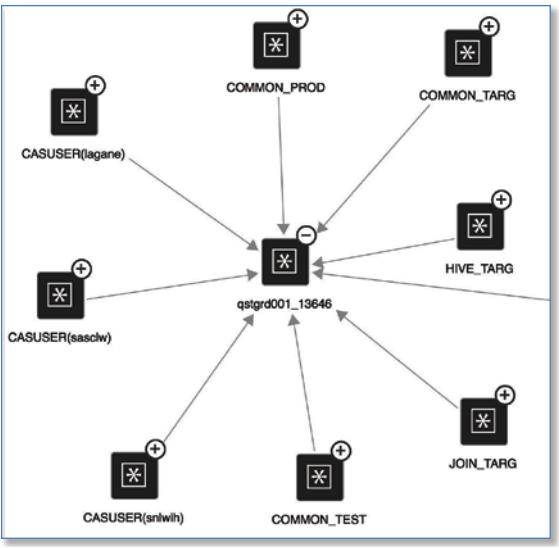
SAS® Data Preparation

분석용 데이터를 신속하게 준비할 수 있는 셀프 서비스 포인트-앤-클릭(Point-and-Click) 환경



SAS® Data Preparation은 어떤 솔루션인가?

SAS Data Preparation은 사용자가 인터랙티브 셀프 서비스 환경에서 리포트 생성 또는 분석을 위해 데이터에 엑세스하고 전처리과정을 수행할 수 있도록 해줍니다.

SAS® Data Preparation이 중요한 이유는?

SAS Data Preparation은 리포트 생성 및 분석용 데이터를 준비하기 위한 예비 작업 시간을 줄여줍니다. SAS Data Preparation의 직관적인 인터페이스는 중요한 기능들을 포인트-앤-클릭 방식으로 제공하므로 특별한 코딩이나 SQL기술이 필요하지 않습니다. 사용자들은 분석 프로세싱과 관련한 활동의 일부로써 완벽하게 정의되어있는 단순한 데이터 준비 작업을 이용하여 데이터 분석에 더 많은 시간을 할애하고 데이터 준비 시간은 줄일 수 있습니다.

SAS® Data Preparation은 누구를 위한 솔루션인가?

SAS Analytics Pro는 비즈니스 분석가, 일반 데이터 사이언티스트와 기타 비전문 사용자를 위해 설계되었습니다. 데이터 사이언티스트와 IT 부서는 동일한 인터페이스를 사용하여 비즈니스 분석가가 재사용할 수 있는 모델을 만들 수 있습니다.

급박하게 변하는 환경에서 질문에 답하기 위해 기업은 일관적이고 신뢰할 수 있는 분석용 데이터에 빠르게 접근할 수 있어야 합니다. 그렇지 않다면 시장과 고객의 요구사항에 빠르게 대응하기 어렵습니다. 미가공 데이터는 오류나 중복 또는 현실을 반영하지 못하거나 병합 시 필요한 식별자가 없는 경우가 많습니다. 분석가는 이러한 데이터 전처리 과정으로 최대 80%나 되는 시간을 허비하기도 합니다. 더욱이 비전문 사용자는 분석용 데이터를 준비하기 위해 데이터를 이동하고 변환하는 기술이 부족합니다. 대부분의 IT 부서는 비즈니스 사용자에게 데이터를 제공하는 데 시간을 빼앗기기 때문에 보다 전략적 활동에 집중할 수 있는 시간이 부족합니다. 또한, 비즈니스 사용자는 분석을 위한 전처리된 데이터를 얻기 위해 IT 부서가 데이터를 생성할 때까지 기다려야 합니다.

SAS® Viya®를 기반으로 한 SAS Data Preparation[1]의 직관적이고 시각적인 인터페이스를 통해 비즈니스 사용자들은 코딩이나 IT 부서의 지원 없이도 신속하게 데이터를 준비할 수 있습니다. SAS Data Preparation은 빠른 인-메모리 분산 환경에서 실행됩니다. 이를 통해 IT 부서는 데이터 제공과 같은 작업으로부터 벗어나고 비즈니스 분석가와 데이터 사이언티스트는 더 빠르게 비즈니스 인사이트를 도출할 수 있습니다. SAS Data Preparation의 인터페이스는 스케줄링이 가능한 코드를 자동으로 생성하여 소스 시스템 리프레시(Source System Refresh)를 통해 실시간 정보의 제공을 보장합니다. 또한 템플릿을 정의하고 재사용할 수 있으므로 공유 및 협업을 강화할 수 있습니다.

주요 특징

- **향상된 생산성을 제공하는 셀프 서비스 데이터 준비** 특별한 기술이나 코딩 없이도 데이터의 액세스, 통합 및 전처리과정이 가능하며, 데이터 준비 작업은 다운스트림 분석 및 리포트 생성 업무와 자동으로 통합되어 동일한 방식으로 시각화할 수 있습니다.

- **효율성을 높여주는 재사용성과 협업** 자동으로 생성된 코드와 사전에 정의된 변환을 IT 부서와 공유하고 각각의 소스 코드 업데이트와 함께 실행되도록 설정할 수 있습니다. 또한 데이터 준비 작업을 프로젝트 내에 저장하여 다른 사용자와 공유하고 재사용할 수 있도록 합니다.

- **분석 사용자를 위한 빠른 결과를 도출** 사전에 정의된 변환 및 데이터 정리 기능을 통해 사용자는 보다 쉽게 데이터를 탐색하고 조정하여 더 많은 데이터를 탐색할 수 있습니다. 그리고 인-메모리 분산 프로세싱 및 병렬 I/O를 통해 거의 실시간에 가까운 응답을 제공합니다.

- **TCO 절감** 교육이 거의 필요 없는 소프트웨어를 통해 일상적인 리포트 생성 및 분석 데이터 준비 작업을 위한 시각적 인터랙티브 인터페이스를 기존 리소스에 적용하여 리소스의 활용을 극대화합니다.

[1] SAS Visual Analytics(별매품)는 SAS Data Preparation을 사용하기 위한 필수 제품입니다.

주요 기능

오늘날 비즈니스 분석가들은 사용할 수 있는 데이터의 양이 방대해지고 종류도 다양해지면서 특정 질문에 대한 답을 제공하기 위해 데이터를 큐레이팅하게 되었습니다. 이를 위해 분석가들은 하루에도 수차례 같은 데이터를 다양한 측면에서 분석해야 합니다. IT 부서에서 비즈니스 분석가를 대신해 데이터를 준비한 경우에도 분석가는 여전히 특정 니즈를 위해 데이터를 반복적으로 조사하고 추가적인 준비가 필요합니다.

SAS Data Preparation은 오늘날 비즈니스 분석 전문가들이 간절히 원하는 임시 환경을 제공합니다. 셀프 서비스 데이터 준비를 위해 설계된 단순하고 직관적인 사용자 인터페이스는 비전문 사용자가 분석을 위한 사실상 모든 소스로부터 데이터를 통합 및 정리, 준비할 수 있는 유연성을 제공하여 분석을 더 빠르고 쉽게 수행할 수 있도록 합니다. 메모리 내에서 로딩된 데이터를 여러 사용자가 동시에 확인할 수 있습니다. 데이터 준비 작업은 모두 동일한 직관적 인터페이스에서 다운스트림 리포트 생성 및 분석 프로세싱과 완전히 통합됩니다. 업계에서 검증된 데이터 통합 및 데이터 품질 기능들이 사전 구축되어 있어 데이터 검사 및 수정 시간을 단축할 수 있습니다. 또한 전체 분석 라이프 사이클에 걸쳐 완전하고 일관적인 분석 환경을 경험할 수 있습니다.

사용하기 쉬운 기능들

SAS Data Preparation을 이용하여 데이터를 간편하게 액세스하고 통합, 보고 및 정리할 수 있습니다. 외부 데이터 소스, Hadoop과 같은 빅 데이터 스토어, SAS Viya의 데이터 등을 시각적으로 탐색합니다. 각각의 상황에 맞게 외부 데이터 소스와 연결하여 필요할 때마다 데이터의 큐레이팅을 수행할 수 있습니다. 또한 열(Column) 이름, 데이터 유형, 인코딩, 열 및 행 수와 같은 물리적 메타데이터 정보를 프로파일링하여 데이터로부터 신속하게 인사이트를 얻습니다.

플랫 파일, 관계형 데이터 소스, 소셜 미디어 소스, SAS 데이터 세트, Apache Hadoop, Teradata, CSV 파일, 텍스트 파일 등 다양한 소스로부터 데이터에 액세스할 수 있습니다. 코딩을 선호하는 사용자의 경우 SAS 코드 또는 Python과 같은 타사 코딩 언어로 SAS Data Quality 작업을 할 수 있습니다.

- SAS Viya에서 허용된 내부 소스, 액세스 가능한 외부 데이터 소스 및 인-메모리 데이터 활용 가능
  - SAS Viya의 인-메모리 엔진에서 로딩되었거나 SAS/ACCESS로 등록된 데이터 소스로부터 로딩된 테이블 또는 파일 샘플을 파악하여 작업할 데이터 시각화
  - 외부 데이터 소스로의 연결 또는 외부 데이터 소스 간 연결을 빠르게 생성
  - 열 이름, 데이터 유형, 인코딩, 열 개수, 행 개수와 같은 물리적 메타데이터 정보에 액세스하여 데이터로부터 더 심층적인 인사이트 획득
- 데이터 소스 및 유형:
  - DNFS, HDFS, PATH 기반 파일(CSV, SAS, Excel, delimited)
  - DB2, Hive, Impala, SAS® LASR™, ODBC, Oracle, Postgres, Teradata
  - Twitter, YouTube, Facebook, Google Analytics, Google Drive, Esri 및 로컬 파일로부터 생성된 피드
  - SAS® Cloud Analytic Services (CAS)

속도 및 확장성

고성능 고품질 데이터로부터 탁월한 분석 결과를 얻을 수 있습니다. SAS Data Preparation을 통해 사용자는 배치(Batch) 프로세스를 기다릴 필요 없이 실시간에 가까운 속도로 인터랙티브하게 데이터 전처리과정을 수행할 수 있습니다. 이러한 기능들은 동시에 로딩되어 메모리 내에서 처리됩니다. 일부 소스의 경우 데이터가 있는 장소에서 처리되므로 SAS 코드의 실행 속도를 높이고 데이터 이동을 최소화하며 신속한 응답을 제공합니다.

셀프 서비스 데이터 준비를 위한 시각적 인터페이스

비즈니스 분석가와 데이터 사이언티스트는 마법사 기반 인터페이스를 사용하여 데이터의 액세스, 통합, 확인, 필터링, 결합, 변환, 정리 및 쿼리를 수행할 수 있습니다. 각각의 변환은 사용자에게 데이터 조정 프로세스를 안내하도록 설계되어 있어 단일 데이터 준비 작업이 결과에 어떻게 영향을 미쳤는지를 쉽게 이해할 수 있습니다.

- 자동 생성 코드로부터 데이터 가져오기 작업을 생성하여 통합 스케줄러를 통해 데이터 리프레시(Data Refresh) 수행
- 데이터 익스플로러 가져오기를 작업으로 설정하여 자동 반복 프로세스 설정
- 작업에 대한 시간, 날짜, 빈도 및 주기 지정
- 코드 작성이나 ETL 툴에 대한 경험 없이 원하는 데이터 소스를 선택하여 데이터를 메모리로 병렬 로딩 가능[2]
- 데이터를 제공하기 전에 행 필터링(Row Filtering) 또는 열 필터링(Column Filtering)을 수행하여 복사되는 데이터양 축소
- SAS In-DB 애드온을 포함시켜 빅 데이터를 원래의 장소에 유지하고 프로세싱을 Teradata나 Hadoop으로 푸시
- 데이터 준비 프로세스를 안내하는 인터랙티브 시각화 환경에서 데이터 변환, 전처리과정 및 표준화 작업 수행
- 실시간으로 시각화 피드백을 얻어 변환이 어떻게 결과에 영향을 주었는지를 쉽게 이해할 수 있는 SAS Viya의 분산형 인-메모리 프로세싱

다양한 사전 구축 변환

SAS Data Preparation에는 열 기반, 행 기반, 코드 기반, 멀티 입력 기반 변환과 같은 사전 구축된 변환이 포함되어 있습니다. 이러한 변환을 통해 보다 쉽게 데이터의 필터링, 전처리과정, 문제 해결 및 표준화 작업을 수행할 수 있습니다.

[2] Twitter, YouTube, Facebook, Google Analytics, Esri 등의 데이터 소스에서는 데이터를 받을 수만 있고 돌려보낼 수는 없습니다.

열 기반 변환

- 열 기반 변환을 사용하여 환경 설정 없이도 데이터의 표준화, 문제 해결 및 세이핑 수행
  - 케이스 변경, 열 변환, 이름 변경, 이동, 나누기, 공백제거/공란제거, 사용자 정의 계산

행 기반 변환

- 행 기반 변환을 사용하여 데이터 전처리과정 수행
- 분석 및 보고 작업용 데이터를 준비하기 위해 전치 변환을 사용한 분석 기반 테이블 생성
- 단순하거나 복잡한 필터를 생성하여 불필요한 데이터 제거

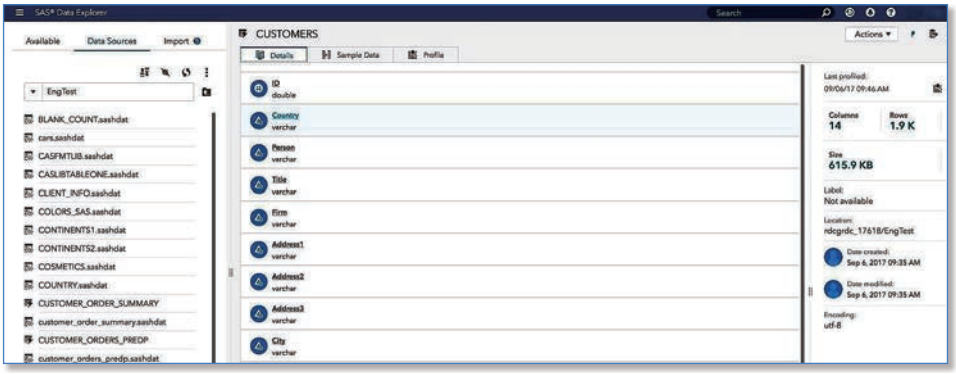


그림 1. 여러 소스로부터 액세스된 데이터를 탐색합니다.

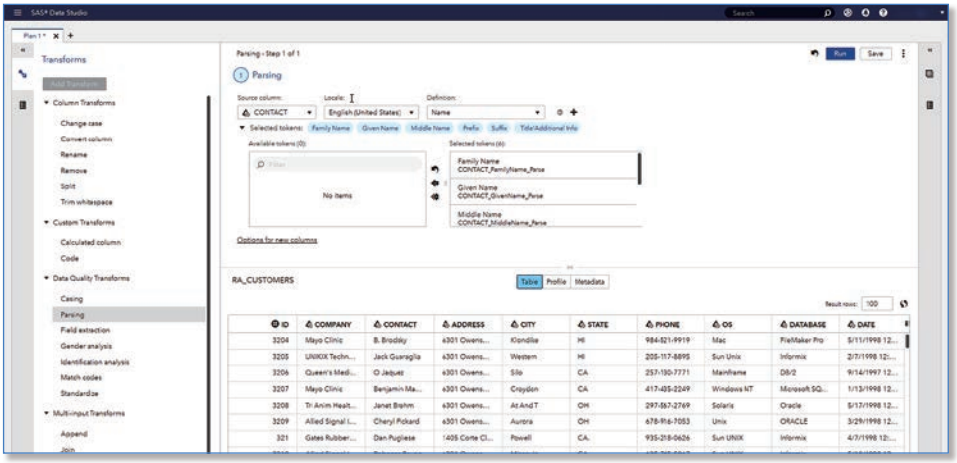


그림 2. 손쉽게 데이터를 변환하고 블렌딩할 수 있는 직관적인 셀프 서비스 환경

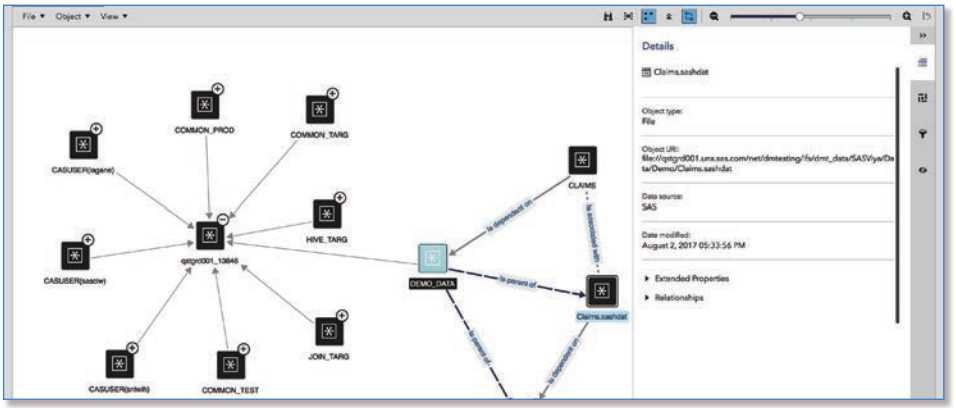


그림 3. 객체 관계도는 다양한 객체 사이의 관계를 보여줍니다.

코드 기반 변환

- 사용자 정의 코드를 작성하여 데이터의 변환, 데이터 전처리, 문제 해결 및 표준화 수행
- 간단한 수식을 작성하여 계산된 열을 생성하고 고급 코드를 작성하거나 코드 정보를 재사용하여 더 유연한 변환 실행
- 다른 사용자가 사용자 정의한 코드를 사용해서 best practice 및 협업 생산성을 공유

멀티 입력 기반 변환

- 멀티 입력 기반 변환을 사용하여 데이터 전처리 수행
- SQL이나 SAS에 대한 지식이 없어도 지정된 인터페이스를 통해 하나 이상의 데이터 세트를 손쉽게 전처리 가능. 데이터 추가, 조인(Join), 전치(Transpose) 및 프로파일링 수행 가능

내장 데이터 품질 기능

SAS Data Preparation에는 분석 레디(Analytics-ready) 데이터 작성을 도와주는 SAS Data Quality 기능이 포함되어 있습니다. 이러한 기능에는 프로파일링, 케이싱(Casing), 표준화, 파싱(Parsing), 식별, 분석 등이 포함됩니다. 사용자는 열 기반 및 테이블 기반 기초 및 고급 프로필 지표를 생성하여 데이터 품질 문제를 찾아내고 데이터 자체에 대한 인사이트를 얻을 수 있습니다. 데이터 품질 및 기타 데이터 준비 작업은 Python과 같은 SAS가 아닌 코딩 인터페이스에서도 액세스할 수 있습니다.[3]

- 데이터를 재처리하고 추가적인 인사이트를 제공하는 저장위치별 또는 데이터분석단위별 파싱(Parsing) 및 필드 추출 정의
- 추출 변환을 통해 지정된 열에서 연락처 정보(이름, 성별, 분야, 패턴, 신원, 이메일, 전화번호 등)를 식별 및 추출
- 파싱(Parsing) 기능을 사용하여 지정된 열의 데이터를 부분열로 토큰화 (예: 성명을 경칭, 성, 이름으로 토큰화 가능)
- 서로 다른 데이터 소스를 연결하는 매치 코드로부터 고유식별자 도출
- 저장위치별 및 분석단위별로 데이터를 표준화하여 케이싱(Casing)과 같이 데이터를 일반 포맷으로 변환
- 로컬별 규칙을 통해 열 데이터를 분석하여 성별 또는 컨텍스트 확인
- 식별 분석을 통해 데이터를 분석하고 특히 데이터 또는 데이터 소스가 잘 알려지지 않은 경우 유용한 데이터 컨텍스트를 확인
- 로컬별 규칙을 사용하는 성별 분석을 통해 이름의 성별을 확인하여 데이터를 간편하게 필터링 또는 분할
- 데이터를 프로파일링하여 열 및 테이블에 기반하는 기초 및 고급 프로필 지표 생성
- 테이블 수준 프로필 지표를 사용하여 데이터 품질 이슈를 찾아내고 데이터 자체에 대한 보다 심층적인 인사이트 제공
- 열 수준 프로필 지표를 위해 각 열로 드릴-다운(Drill-Down)하여 패턴 분포 및 빈도 분포 결과에 대한 그래프를 확인하고 숨겨진 인사이트 확인
- 앞서 열거된 다양한 데이터 유형/소스를 사용하여 Twitter, Facebook, Google Analytics 또는 YouTube로부터의 데이터 프로파일링

데이터 거버넌스 및 관계도

SAS Data Preparation을 통해 사용자는 데이터 소스, 데이터 객체, 데이터에 적용한 기능들을 탐색하여 파이프라인 활동을 손쉽게 추적할 수 있습니다.

- 액세스 가능 데이터 소스, 데이터 객체, 작업 간 관계 확인
- 객체 간 존재하는 관계를 시각화하거나 보다 쉽게 데이터 출처를 파악하고 그 프로세싱을 추적할 수 있는 관계 그래프
- 시스템 및 작업 수준 프로세스를 위한 통합 모니터링 기능
- 얼마나 많은 프로세스가 얼마나 오래 진행 중이고 누가 그 프로세스를 실행하는지에 대한 인사이트 제공
- 작업 상태(실행, 성공, 실패, 대기, 취소)를 기반으로 간단하게 모든 시스템 작업 필터링
- 작업 에러 로그에 액세스하여 근본적인 원인 분석 및 문제 해결 지원

협업, 재사용 및 자동화

SAS Data Preparation을 통해 사용자는 특정 분석에 대한 데이터를 준비하고 변환을 저장하고 공유하여 다른 사용자가 나중에 재사용할 수 있는 기능을 제공합니다. 템플릿을 포인트-앤-클릭(Point-and-Click) 인터페이스 또는 코딩 환경에서 정의하여 다른 사용자가 재사용할 수 있는 최상의 조건을 정의합니다. 또한 템플릿 코드를 IT 프로세싱의 일부로 포함하여 새로고침을 통해 준비된 데이터가 현재 상태를 보여주도록 유지합니다.

- 하나 이상의 데이터 소스에 적용되는 변환 규칙 세트로 구성된 데이터 준비 계획(템플릿)을 이용하여 생산성(데이터 준비 시간 단축) 향상
- 템플릿을 다른 데이터 세트에 재사용하여 해당 데이터가 기업의 데이터 표준 및 정책을 준수하며 일관성 있게 변환되도록 지원
- SAS Viya 프로젝트에 사용되는 프로젝트 허브를 통해 팀 기반 협업 수행. 누가 무엇을 언제 했는지 확인하고 다른 팀 구성원들과 소통할 수 있는 프로젝트 활동 피드

[3] SAS에 대한 타사 인터페이스는 GitHub에서 다운로드할 수 있습니다.

더 자세한 내용은 sas.com/korea/data-preparation에서 확인하실 수 있습니다.

sas
THE POWER TO KNOW.