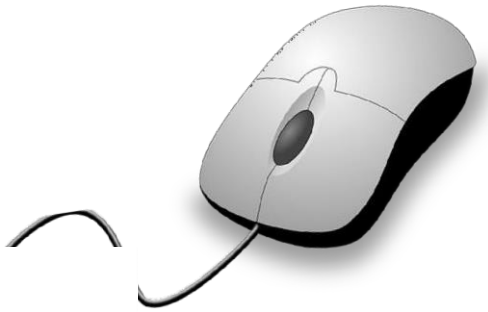
공개SW 솔루션 설치 & 활용 가이드

응용SW > BI/OLAP



pentaho 제대로 배워보자

How to Use Open Source Software

Open Source Software Installation & Application Guide

OPENUP

오픈소스 소프트웨어 통합지원센터

Open Source Software Support Center

CONTENTS

1. 개요

소개	• 다양한 데이터 배포 위한 엔터프라이즈 급 오픈 소스 기반 데이터 통합 및 분석 플랫폼 • 빅데이터 통합, 데이터 웨어하우징, 워크 플로우 통합, 데이터 과학 적용, 데이터 탐색, 데이터 시각화, 보고서 및 대쉬보드 포함한 분석 플랫폼		
주요기능	• 다양한 데이터 소스들 추출, 변환 및 적재할 수 있는 데이터 통합 기능과 머신러닝 기반의 Advanced Analytics • Business Intelligence 및 시각화 • 스케줄링 통한 모델 자동 업데이트 및 챔피언쉽 모델 챌린지 기능으로 최적화 모델 선정		
대분류	• 응용 SW	**소분류**	• BI/OLAP
라이선스 형태	• Apache License V2.0	**사전설치 솔루션**	• N/A
실행 하드웨어	• 서버플랫폼 : Window, 리눅스 등 • CPU : 4 Core 이상, 램 8GB 이상 • 20GB 이상의 디스크 공간 (최소 10GB)	**버전**	• 8.1 (2018년 10월 기준)
특징	• 오픈 소스 기반(Enterprise Edition의 경우 라이선스 비용 발생) • 데이터 수집부터 추출, 변형, 적재, 분석, 시각화까지 원스탑으로 수행 가능한 분석 플랫폼 • 한글화 지원		
보안취약점	• 취약점 ID : CVE-2015-6940 • 심각도 : 5.0 MEDIUM(V2) • 취약점 설명 : Pentaho 버전 5.2.x GA BA Suite 및 PDI는 구성 파일에 대한 인증되지 않은 액세스 허용 • 대응방안 : 사용중인 DI 및 BI 플랫폼 버전에 해당하는 .jar 파일 업데이트 • 참고 경로 : https://www.securityfocus.com/archive/1/536477/100/0/threaded		
개발회사/커뮤니티	• Hitachi Vantara / Pentaho Community		
공식 홈페이지	• https://www.hitachivantara.com/go/pentaho.html		

2. 기능요약

- Pentaho PDI의 주요 기능

Feature ₩ Version	8.1	Feature ₩ Version	8.1
DI-Server (Merged Pentaho Server)		**Enterprise Deployment**	
- Content-Repository	CE	- Job Restartability	EE
- Version control	CE	- Transactional Job Execution	EE
- User/Role Security	CE	- Load Balancing (Transformations)	EE
- Database Security	CE	- Worker Nodes scale out	**EE**
- Purge Utility	CE		
- DI-Scheduling	EE	**Data Science Pack (Data Mining)**	
		R Script Executor	EE
Innovation & Big Data		Weka Forecasting	EE
- Data Services	CE	Weka Scoring	EE
- Carte on YARN	EE	Arff Output	EE
- SDR (Automodelling, Publish)	CE		
- Data Explorer	EE	**Special topics (job entries and steps)**	
		- Agile BI	CE
Monitoring & Auditing		- JMS Support	CE
- PDI Operations Mart	EE	- IBM MQ Support	--
- SNMP Monitoring	EE	- Splunk Support	EE
		- SAP Hana Bulk Loader	EE
Security & Big Data		-MQTT	CE
- AES Password Support	EE		
- Kerberos Support for Hadoop	EE		
- Sentry Support for Hadoop	EE		
- Hadoop HA Support	EE		
- Ranger Support for HDP	EE		
- Knox Security	EE		
- Google BigQuery/Cloud Storage	**EE**		

2. 기능요약

- Pentaho PBA의 주요기능

Feature ₩ Version	8.1
Analyzer	
General Analyzer Functionality	EE
Mondrian OLAP Engine	CE
Geo Plugin	EE
Analyzer JavaScript APIs	EE
Schema Workbench	CE
Aggregation Designer	CE
Dashboards	
Dashboard Designer	EE
Ctools - CDE/CDF	CE
Reporting	
Interactive Reporting (browser based)	EE
Report Designer (pixel perfect / desktop)	CE
Reporting Engine	CE
Pentaho Metadata Editor	CE
Plattform Administration & Auditing	
Audit Reporting (Operations Mart)	EE
Pentaho User Console	CE
Data Source Wizard	CE
Data Source Model Editor	EE
Migration Tool	EE
JDBC Distribution Utility	EE

3. 실행환경

제 품 명	Pentaho DI	Pentaho BA
Version	Pentaho Enterprise Edition	
제품 용도	데이터 추출, 변형, 적재 도구	데이터 분석 도구
서버 플랫폼	Windows Server 2008 R2&2012, CentOS 6&7, RedHat Enterprise 6&7, Ubuntu Server 14.04 LTS & 16.04 LTS, Suse Linux 11(SP3+)	
메타데이터 DB	MySQL 5.6/5.7, Oracle 11.2&12.1, PostgreSQL 9.4 & 9.5+, MSSQL 2012&2014	
클라이언트 플랫폼	Windows 7&10, Ubuntu Desktop 12.04&14.04, OSX 10.10&10.11, iOS 8.x	
Web Browser	Safari 9.x & 10.x, Chrome 53 & 54, Internet Explorer11, Firefox 48 & 49	
Security	Active Directory, LDAP, RDBMS, CAS, Integrated Microsoft Windows Authentication	
JVM	Oracle Java 8	
한글화 지원	메뉴, 메시지, 데이터 처리 등 한글화 지원	
HW 최소 사양	CPU(4Core), Mem(8GB), HDD(20GB)	

4. 설치 및 실행

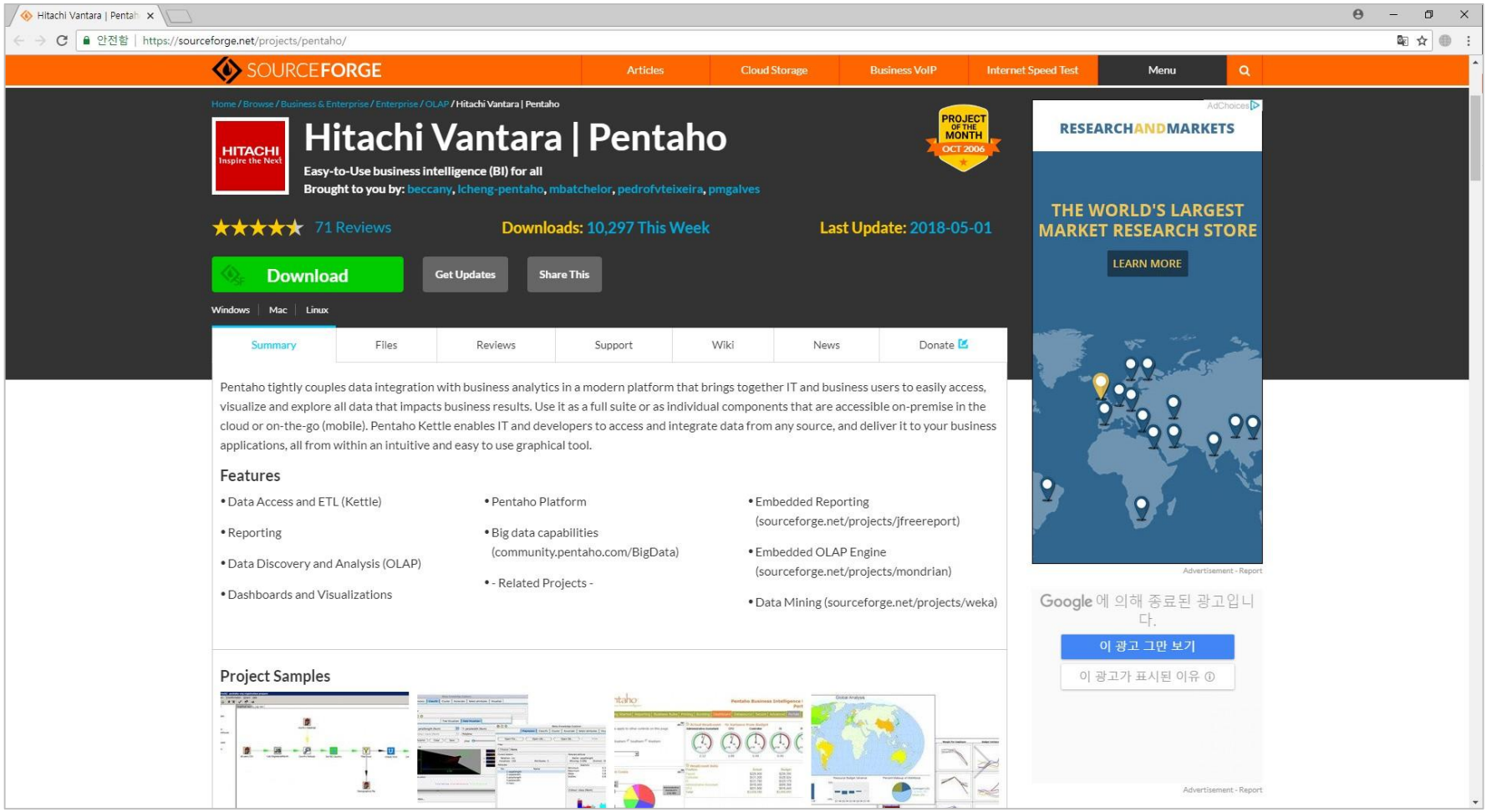
세부 목차

1. 설치 파일 다운로드
2. 실행
3. 설치

4. 설치 및 실행

4.1 설치 파일 다운로드 (CE 버전)

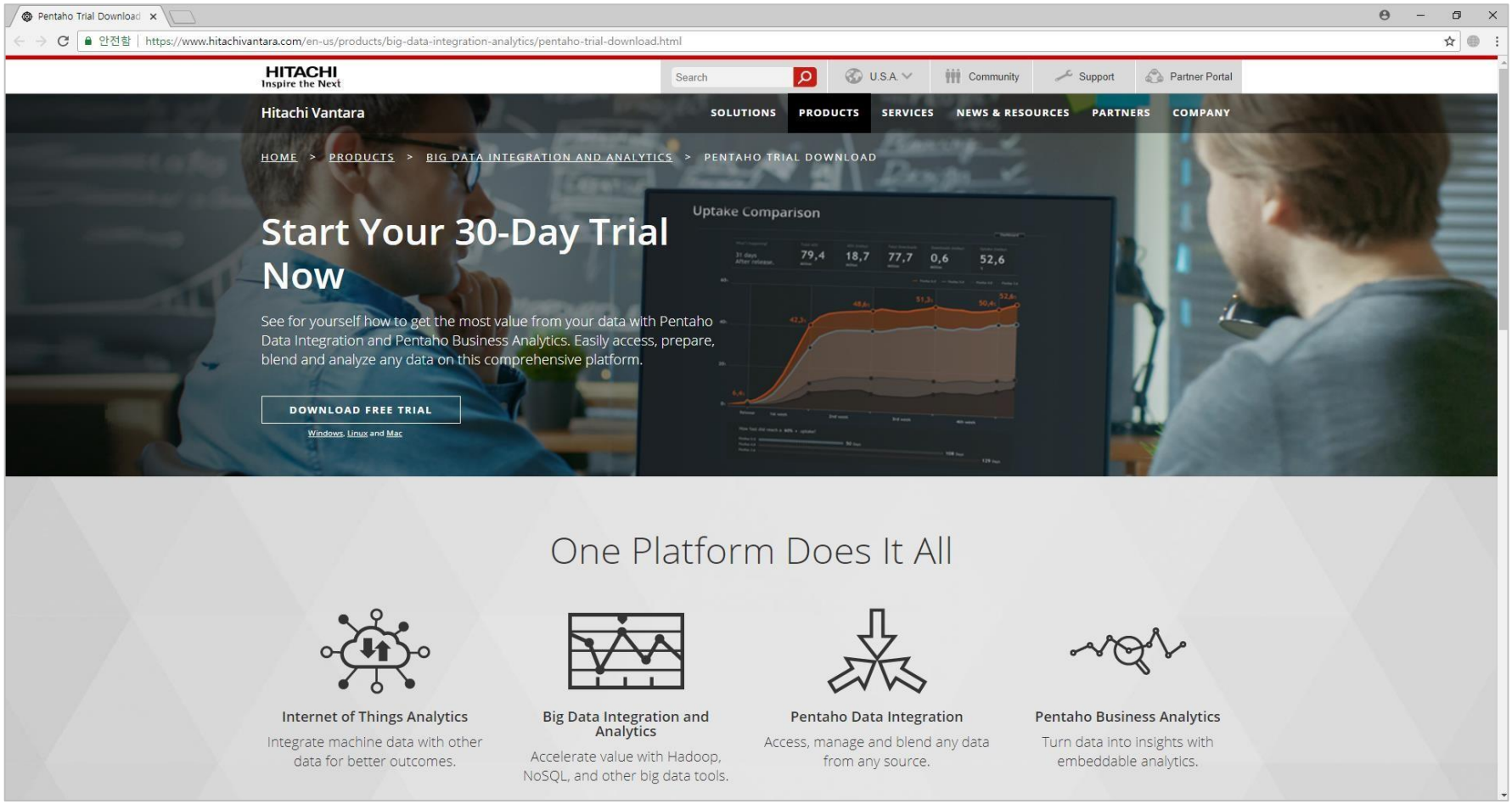
- https://sourceforge.net/projects/pentaho/ 접속 후, Download 클릭



4. 설치 및 실행

4.1 설치 파일 다운로드 (EE 버전)

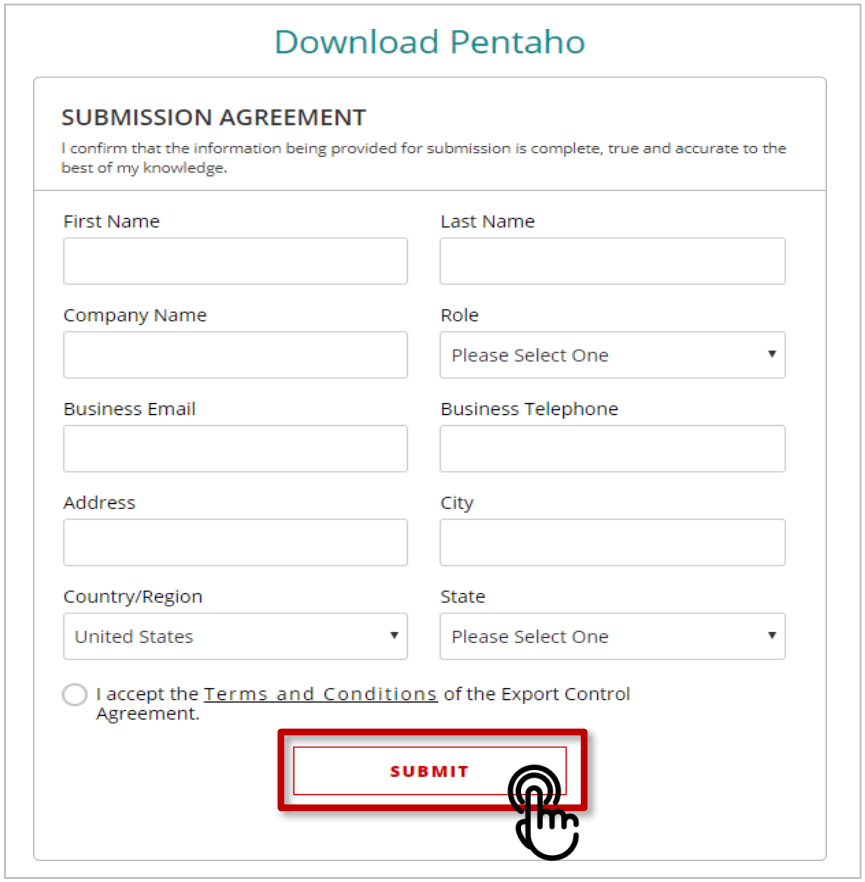
- https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-trial-download.html 접속 후, DOWNLOAD FREE TRIAL 클릭



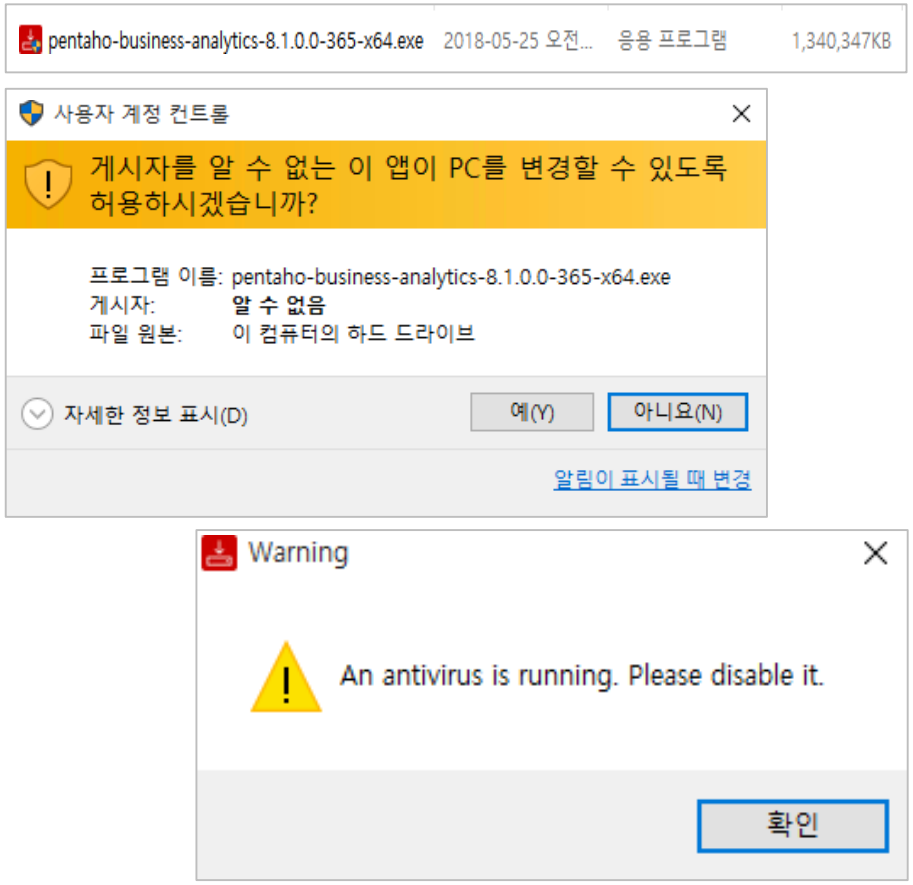
4. 설치 및 실행

4.2 설치(1/5)

- 사용자의 요구하는 기능에 따라 CE 혹은EE 버전 설치하여 사용(4~5page 참고)
- 개인 정보 입력 후 SUBMIT 버튼 클릭



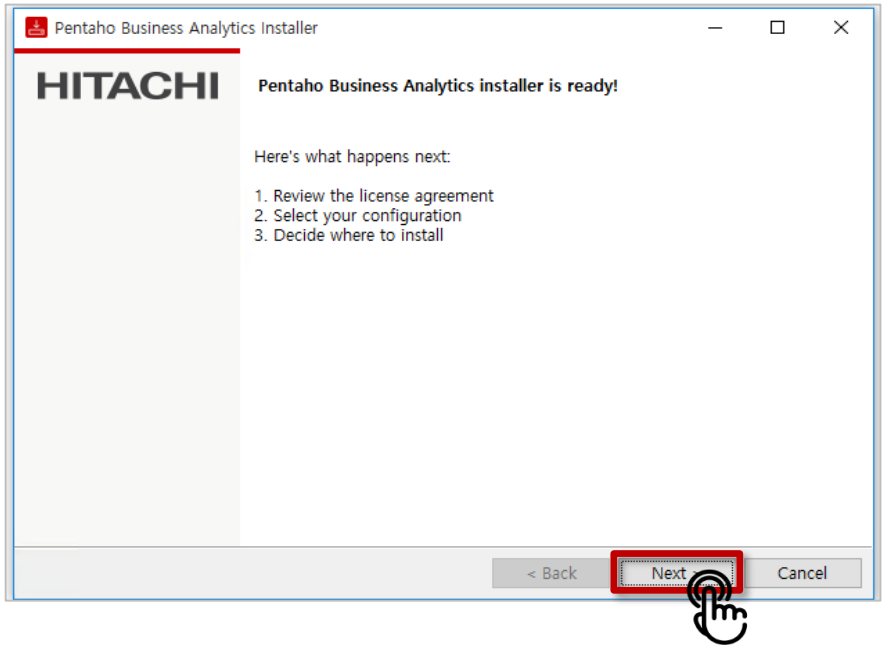
- 다운로드 받은 파일 경로로 이동하여 Pentaho 8.1 설치 파일 실행 후, 예(Y) 클릭, 확인 클릭



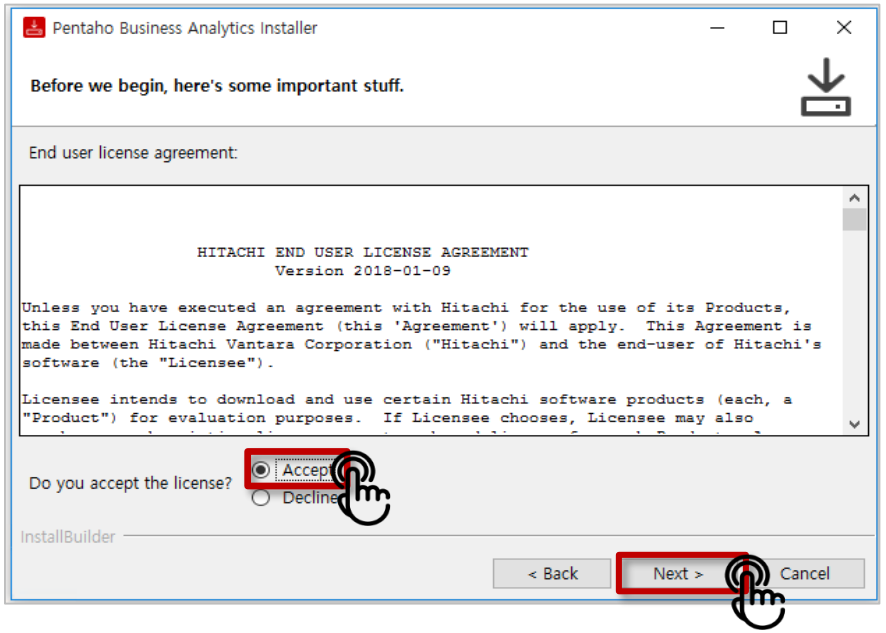
4. 설치 및 실행

4.2 설치(2/5)

- Next 클릭



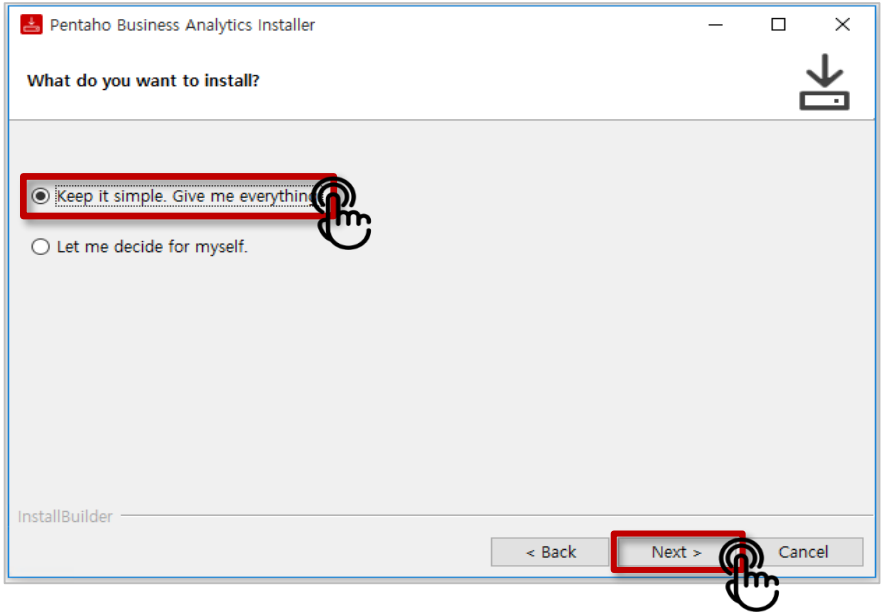
- 'Accept' 선택 후 Next 클릭



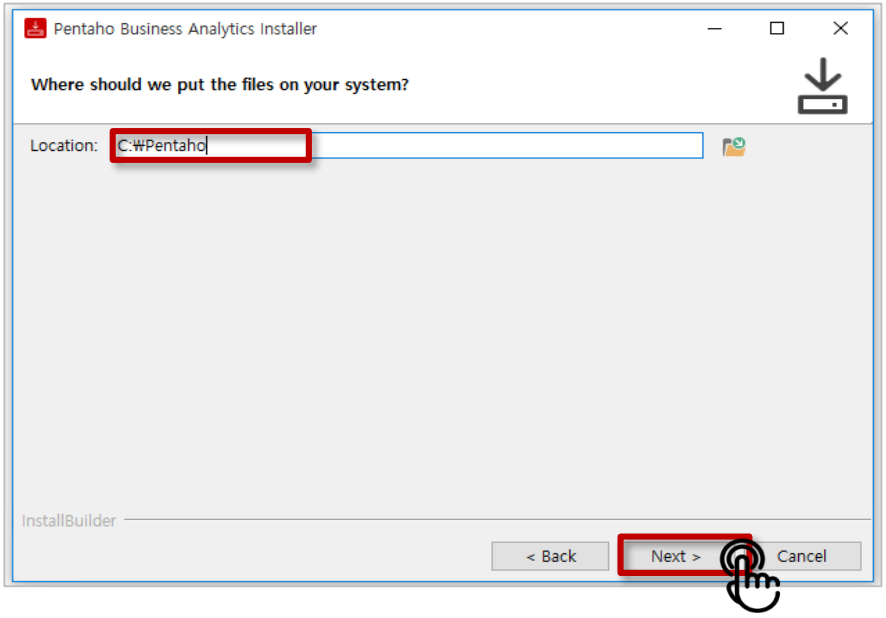
4. 설치 및 실행

4.2 설치(3/5)

- ‘Keep it simple. Give me everything’ 선택 후, Next 클릭



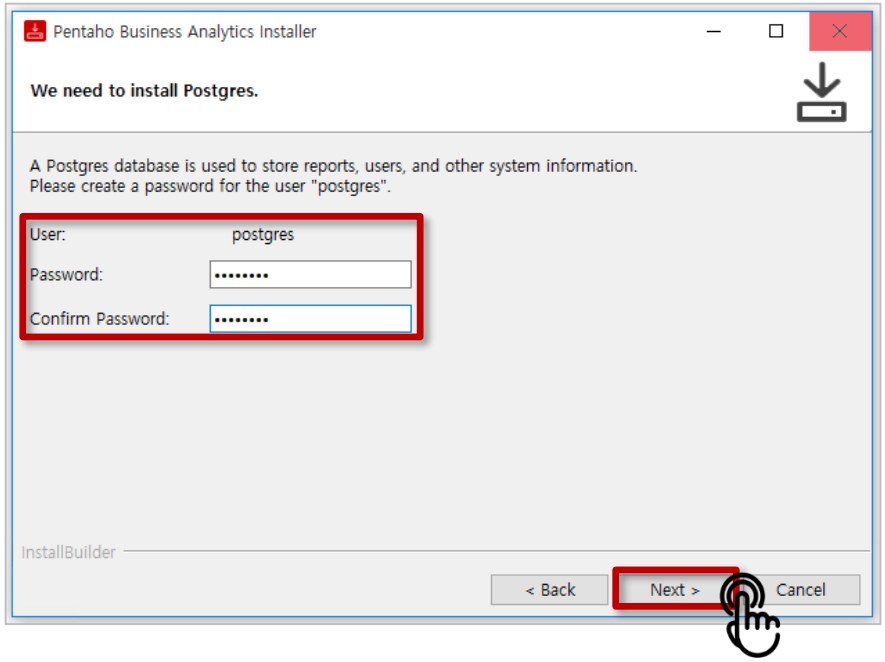
- 설치 경로 설정 후, Next 클릭 (Default 경로 : C:₩Pentaho)



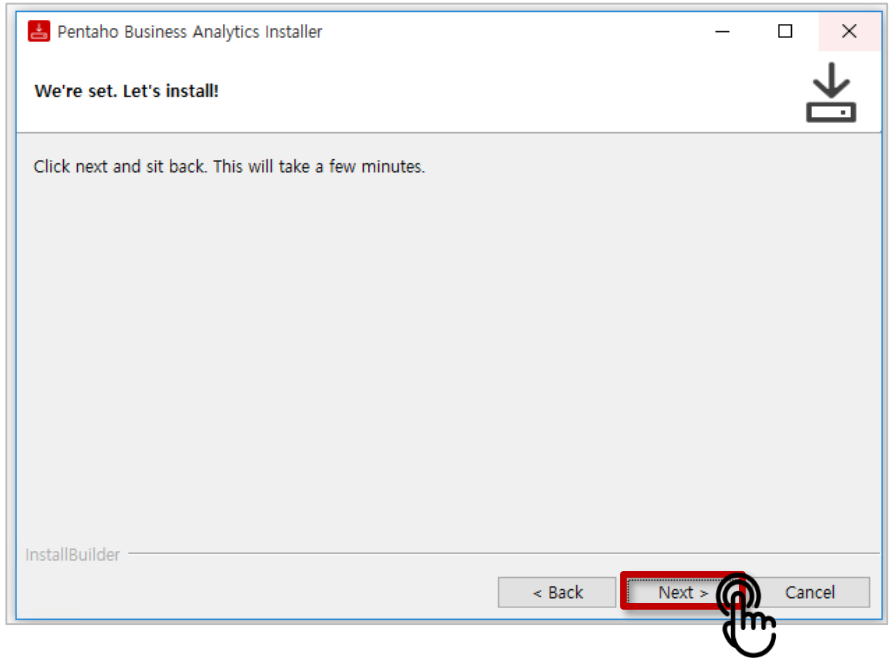
4. 설치 및 실행

4.2 설치(4/5)

- Pentaho 설치 시 함께 설치되는 Postgres에 대한 비밀번호 설정한 후, Next 클릭



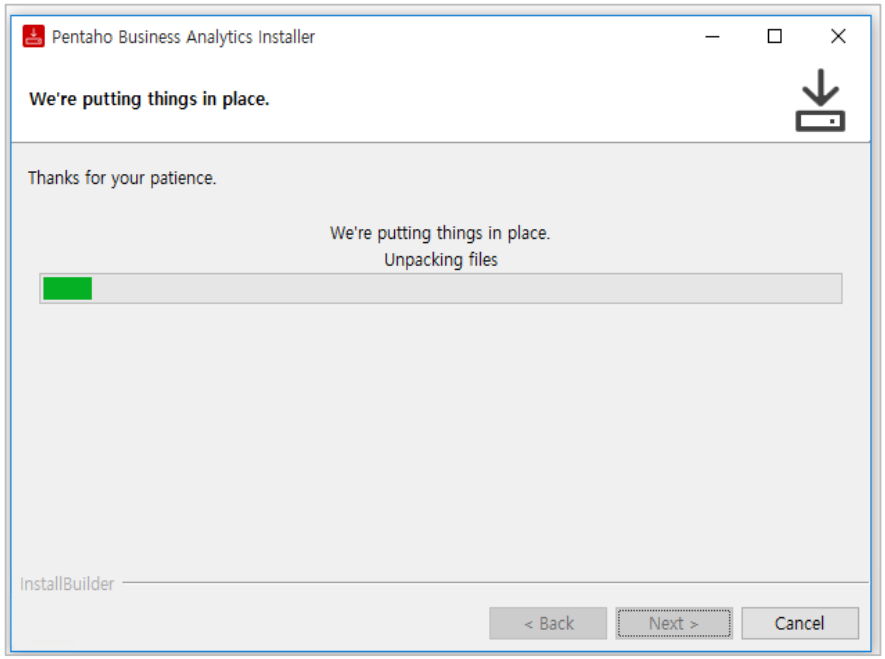
- 설치 준비가 끝나면, Next 클릭



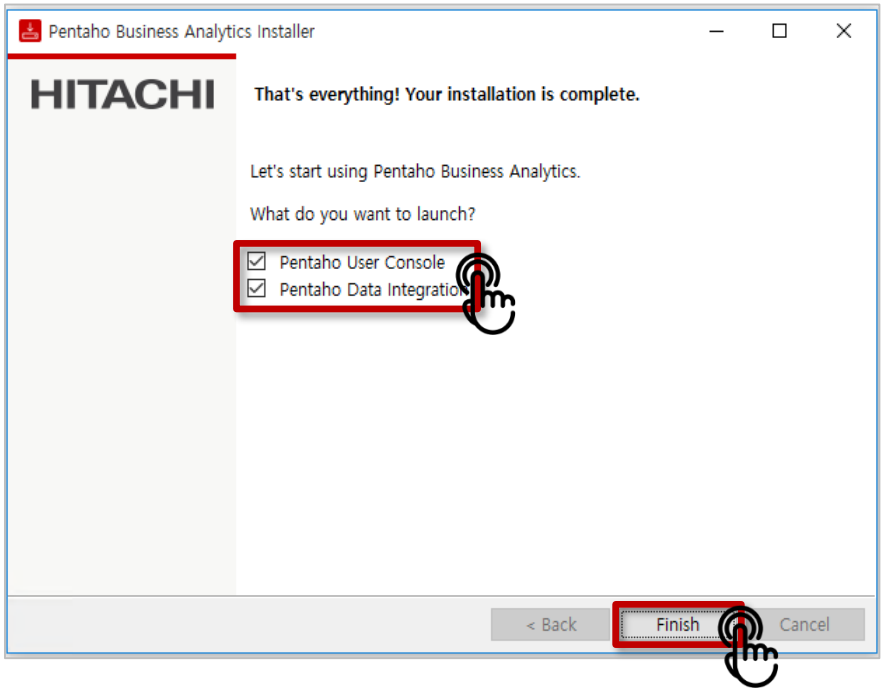
4. 설치 및 실행

4.2 설치(5/5)

- 설치 약 10-20분 소요



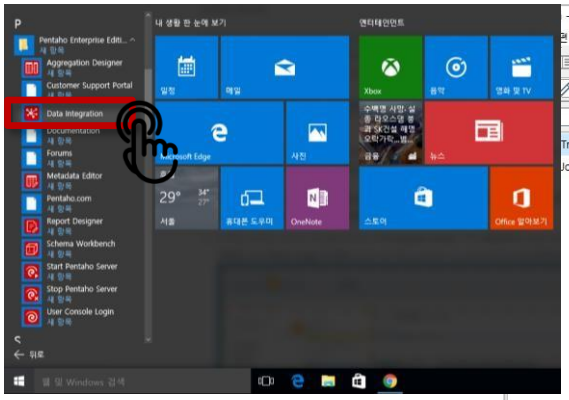
- PUC와 PDI 바로 실행하는 경우 체크박스 선택한 후, Finish 클릭

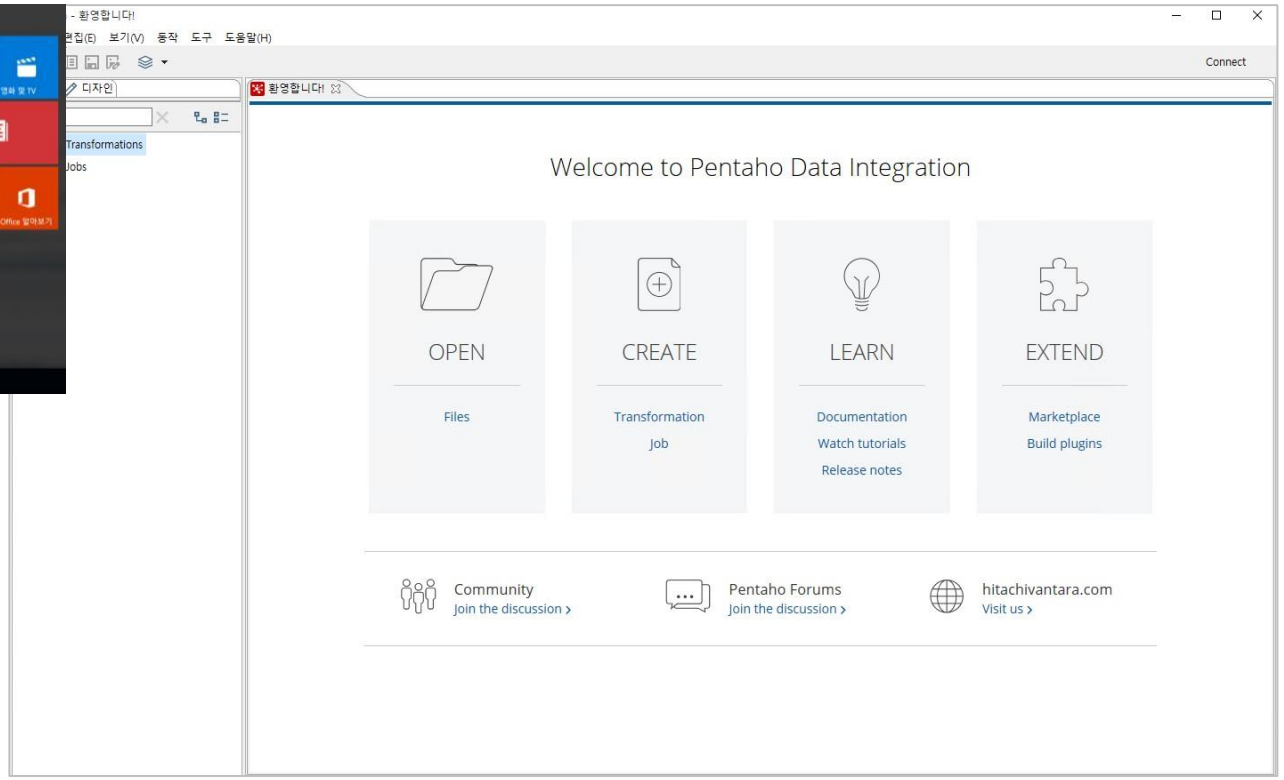


4. 설치 및 실행

4.3 실행(1/2)

- PDI(Pentaho Data Integration) Client 실행 :
  시작 ❼ 모든 앱 ❼ Pentaho Enterprise Edition ❼ Data Integration 클릭 (Windows 10 기준)

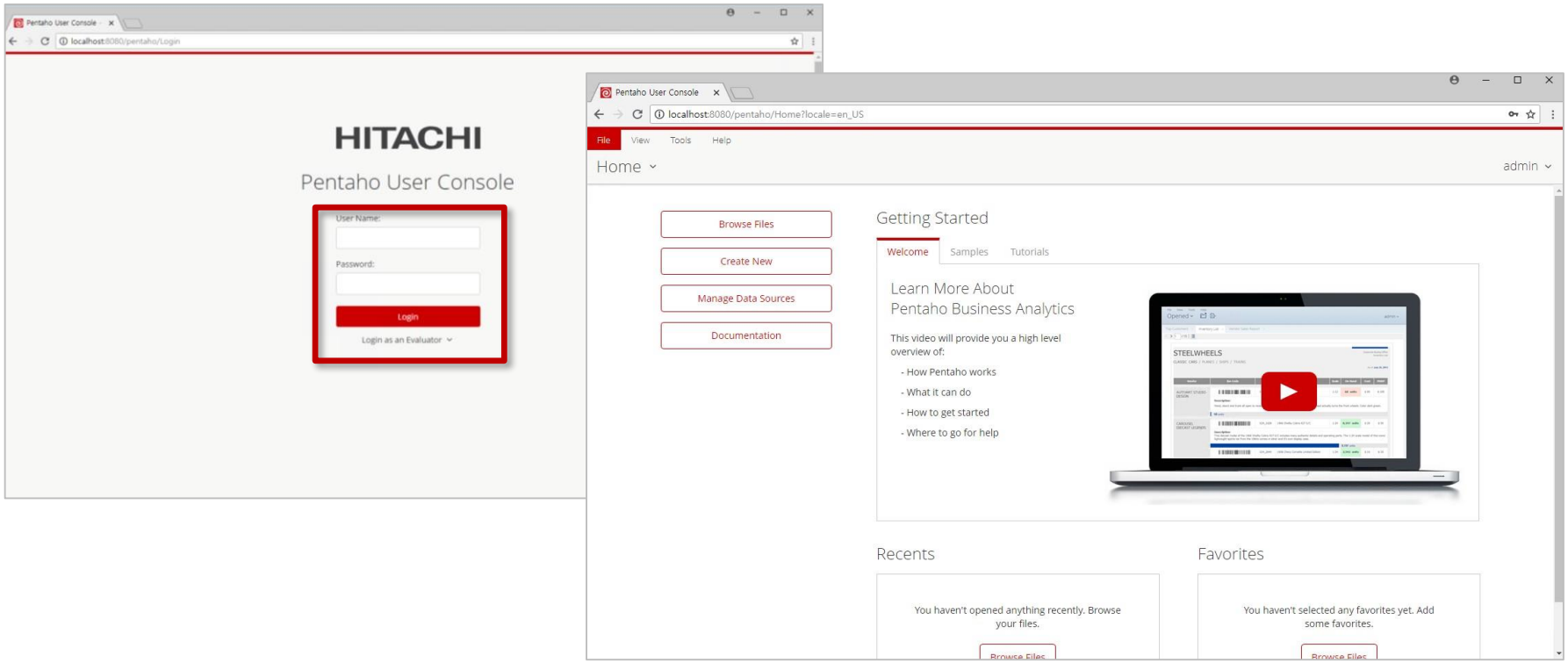




4. 설치 및 실행

4.3 실행(2/2)

- PUC(Pentaho User Console) 실행:
  웹 브라우저에서 http://localhost:8080 입력 후, 로그인 창에 admin/password 입력 후, Login 클릭



5. 기능소개

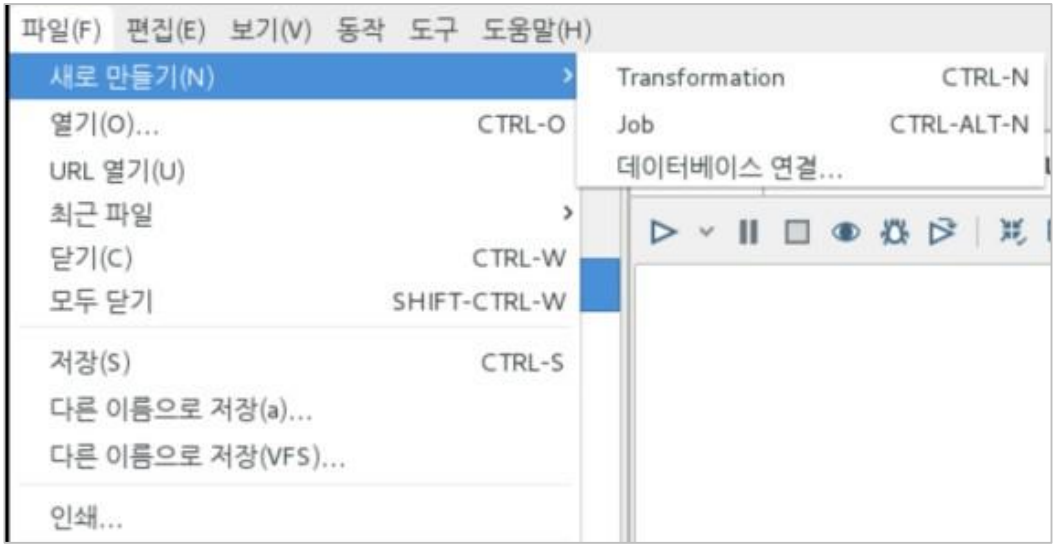
세부 목차

1. PDI Step 활용
2. 데이터 Input
3. Step 연결 및 실행
4. 실행결과 확인
5. DB에 저장
6. Machine learning

5. 기능소개

5.1 PDI Step 활용

- Pentaho Data Integration 열기
- Pentaho Data Integration Trasformation 생성 : 파일 ❼ 새로만들기 ❼ Transformation

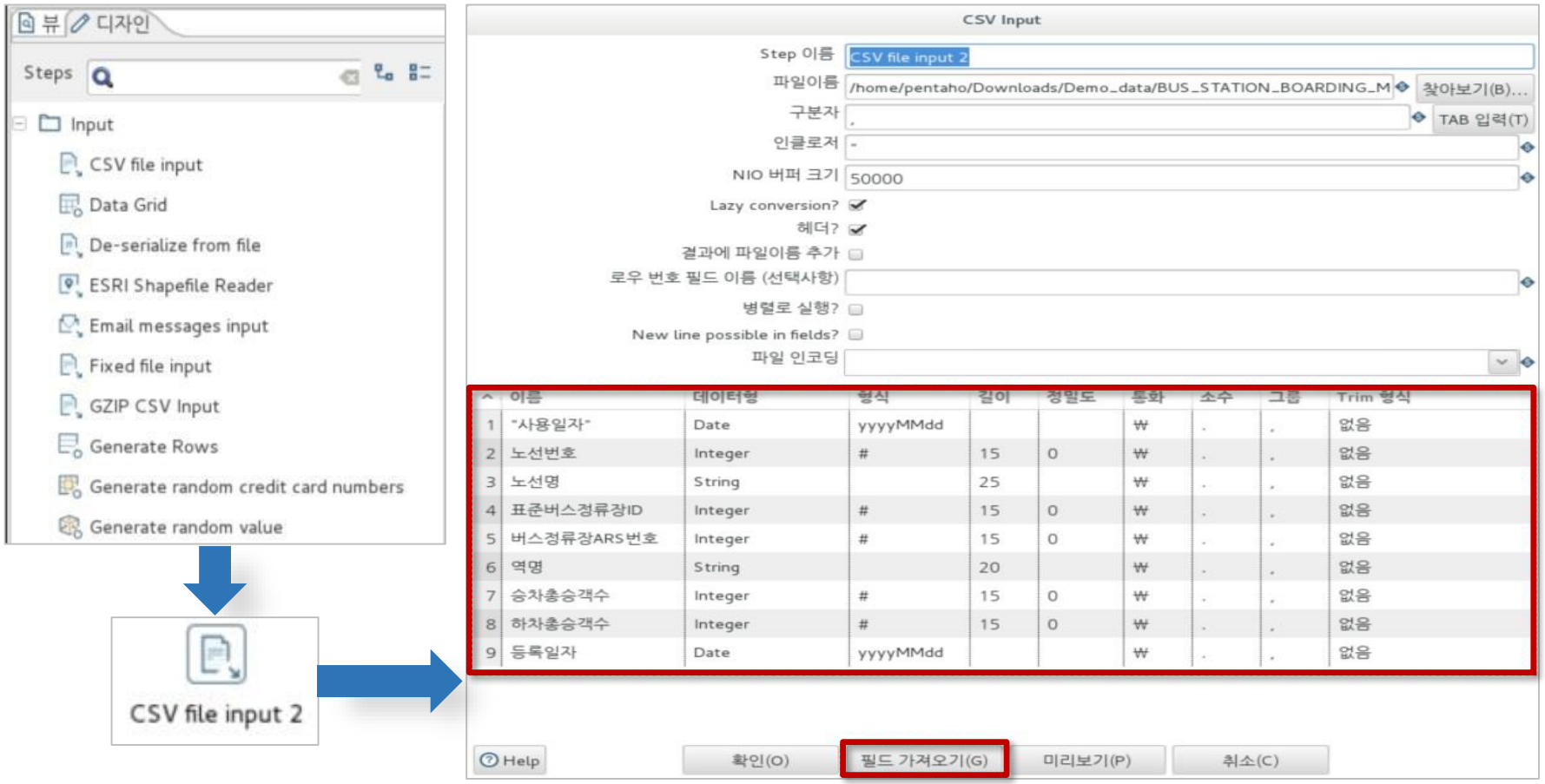


- 파일 저장
  - 하나의 Transformation에 전체 work 담기보다는 필요에 따라 분리하여 Transformation 구성

5. 기능소개

5.2 데이터 Input

- Pentaho Data Integration Step 활용 : 디자인시트에서 원하는 Step 선택 후 드래그

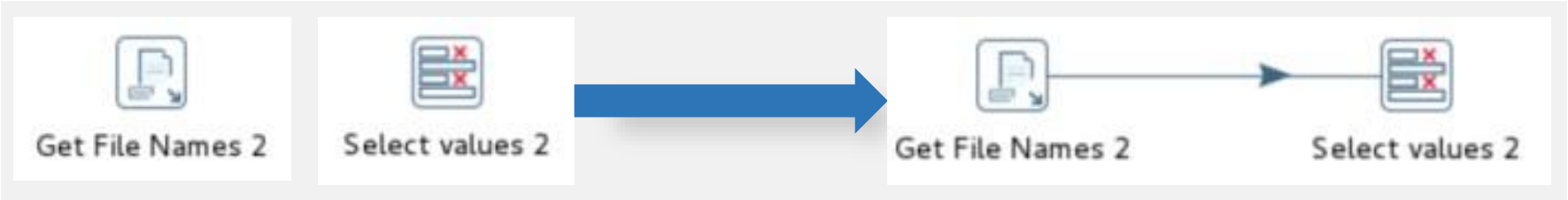


- Step 클릭하여 파일 경로 설정 등 원하는 셋팅 값 입력

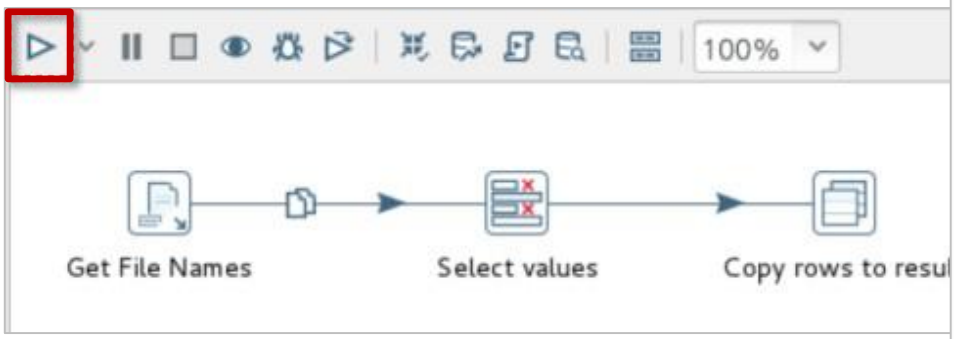
5. 기능소개

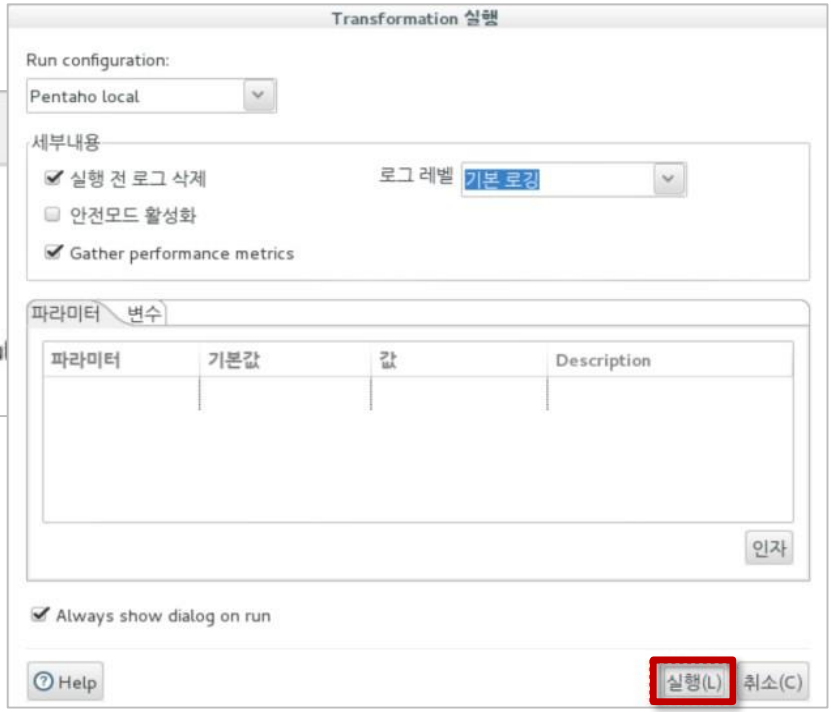
5.3 Step 연결 및 실행

- Hop 연결 : Step 간의 호프 연결하여 워크 플로우 설정



- Transformation 실행

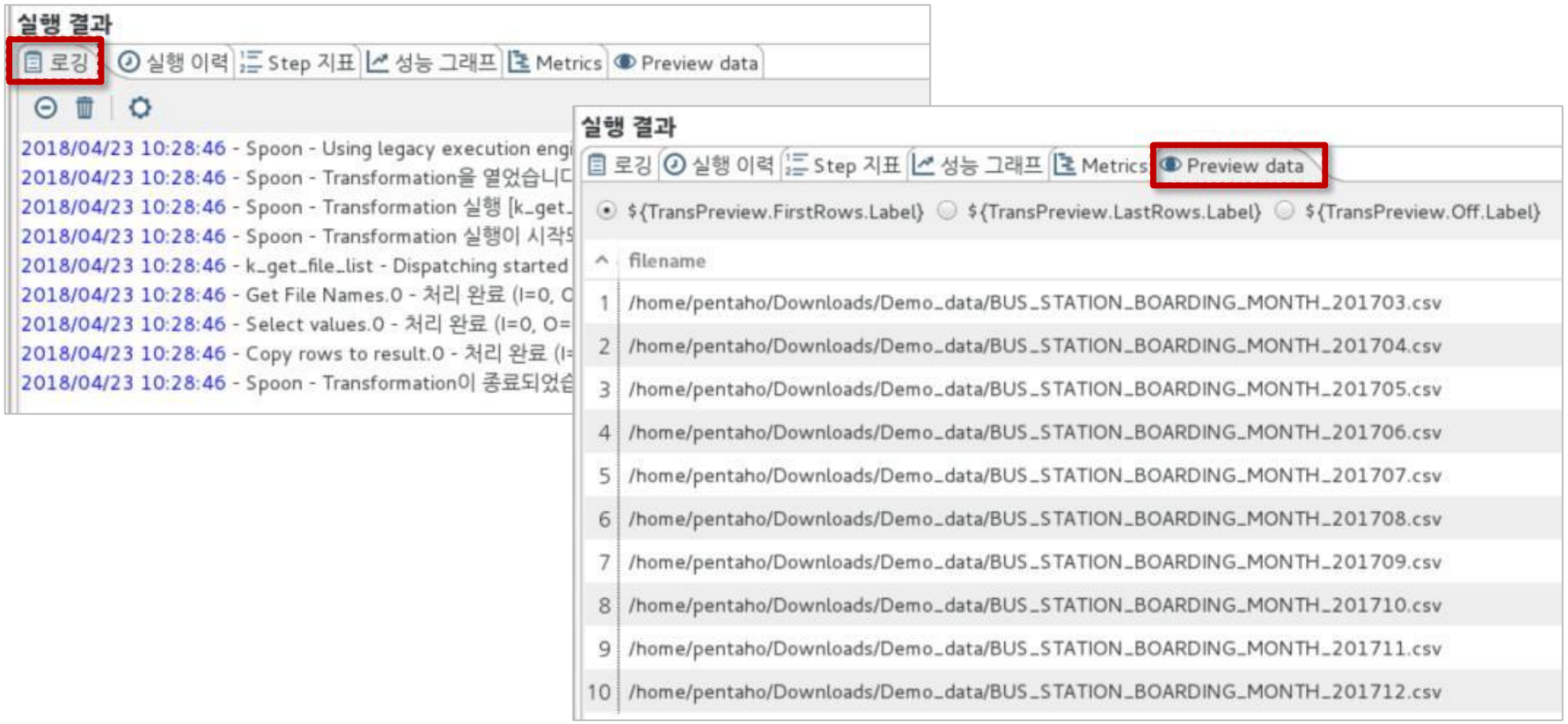




5. 기능소개

5.4 실행 결과 확인

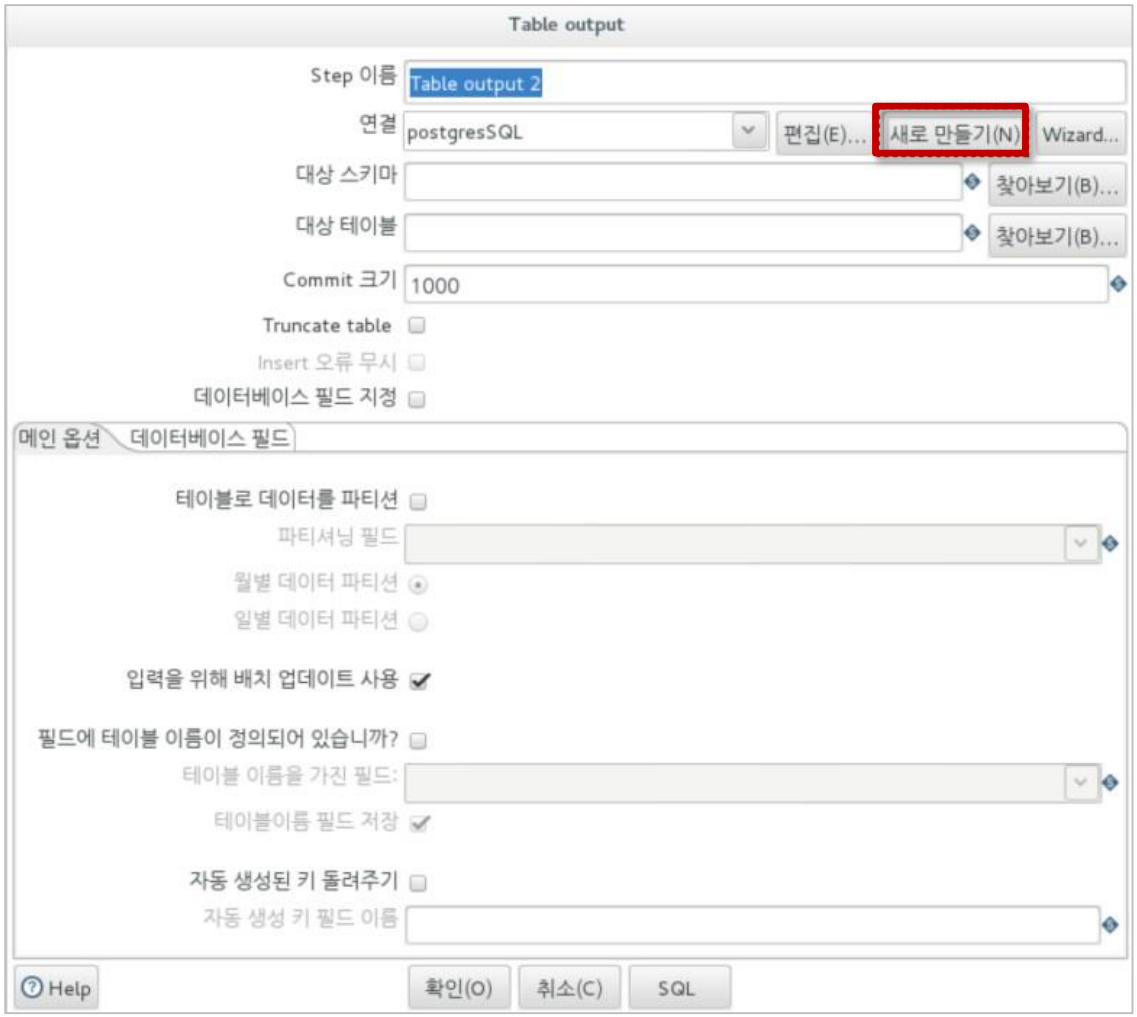
- 실행결과 확인 : 로깅 & Preview data 통하여 실행 결과 및 Transformation 내용 확인



5. 기능소개

5.5 DB에 저장

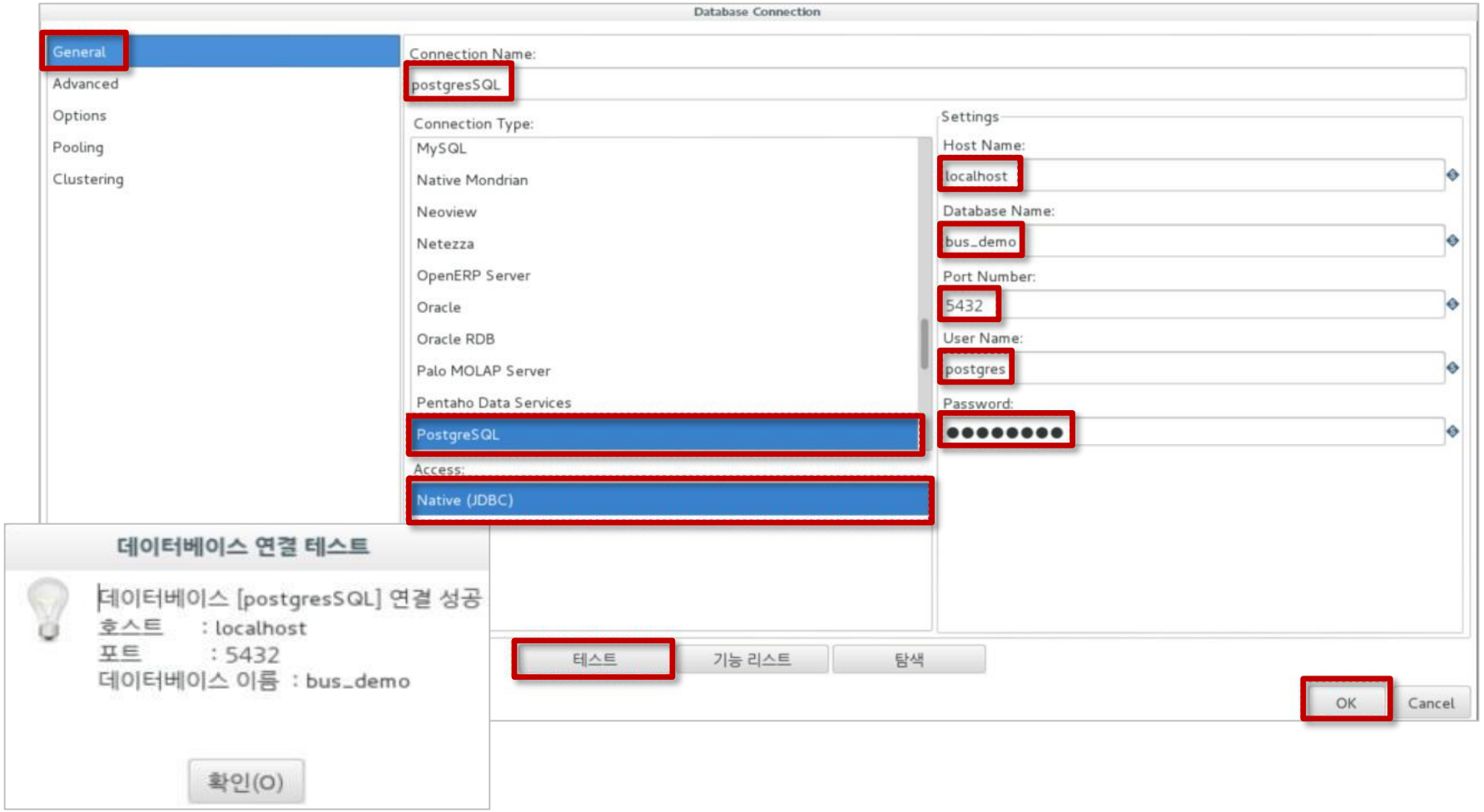
- DB에 저장 : Table output > 새로만들기 > Databse Connection 입력 > 테스트 > 확인



5. 기능소개

5.5 DB에 저장

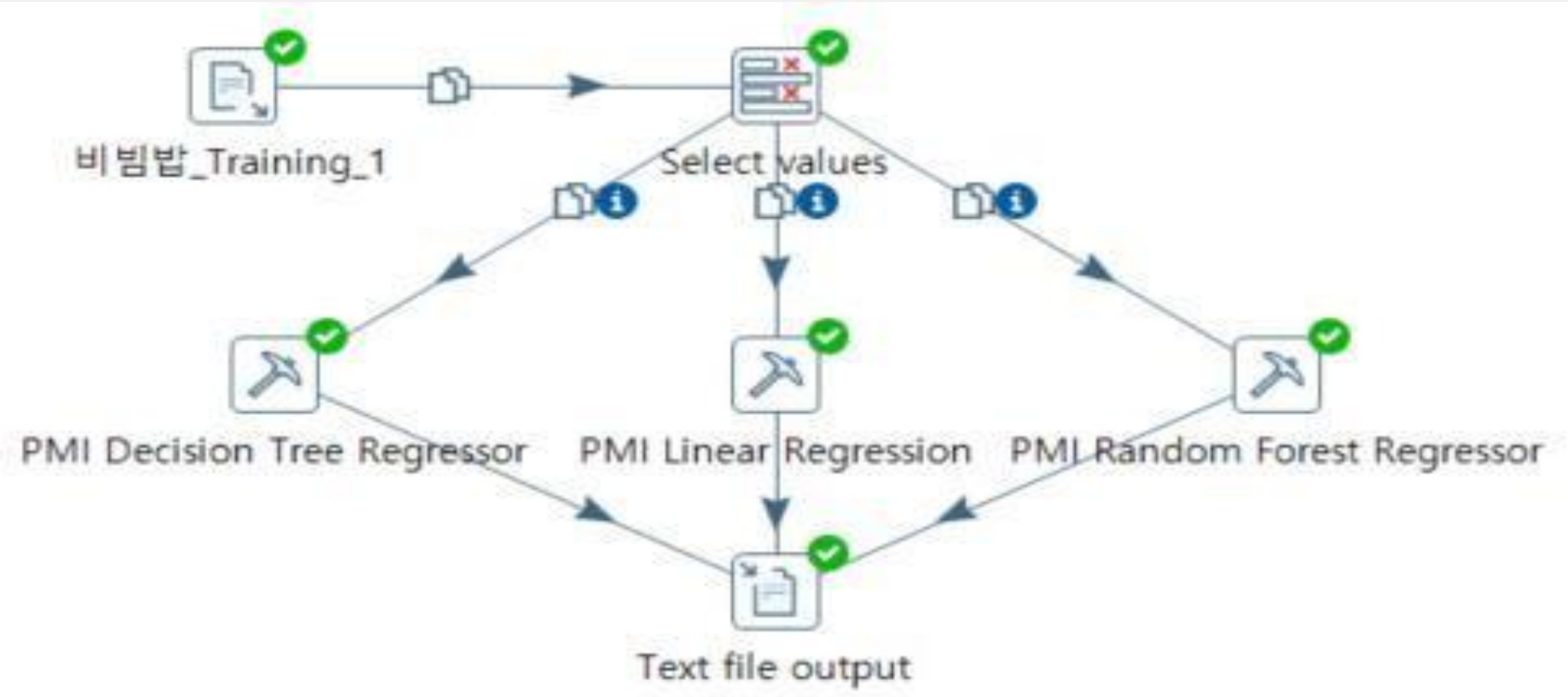
- DB에 저장 : Table output > 새로만들기 > Databse Connection 입력 > 테스트 > 확인



5. 기능소개

5.6 Machine learning (PMI)

- Machine learning step 활용하여 데이터 분석에 활용



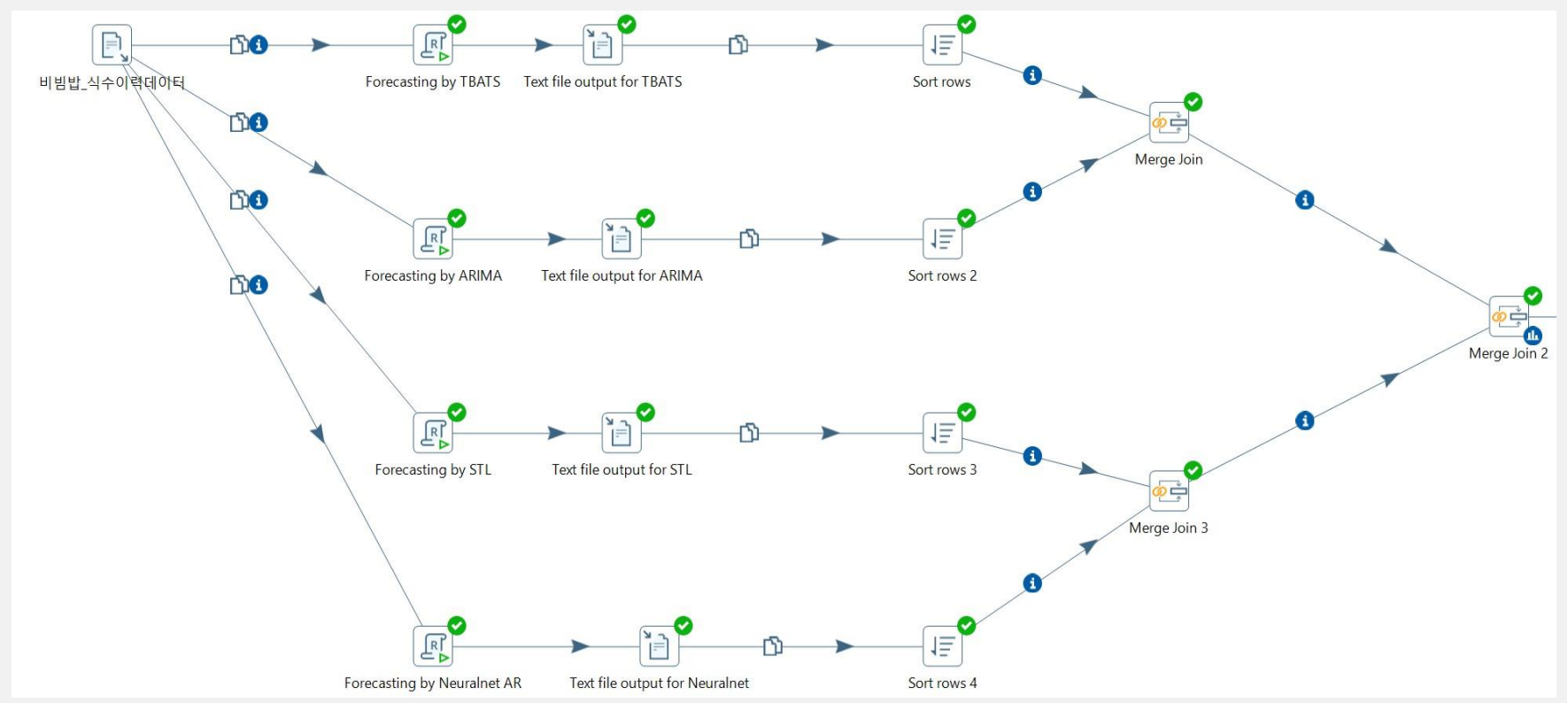
Pentaho 머신러닝구현 workflow

데이터 Input → 변수 선정 → 모형 학습 (ML 모델) → 학습모형 저장 및 업데이트

5. 기능소개

5.6 Machine learning (DataScience Pack)

- R , Python, Weka 등의 오픈소스 스크립트 Step에 플러그인하여 분석에 활용



Pentaho 시계열 예측 workflow

데이터 Input → 시계열 분석 실시 → 모델 예측 결과통합 → 분석DB 저장

6. 활용예제

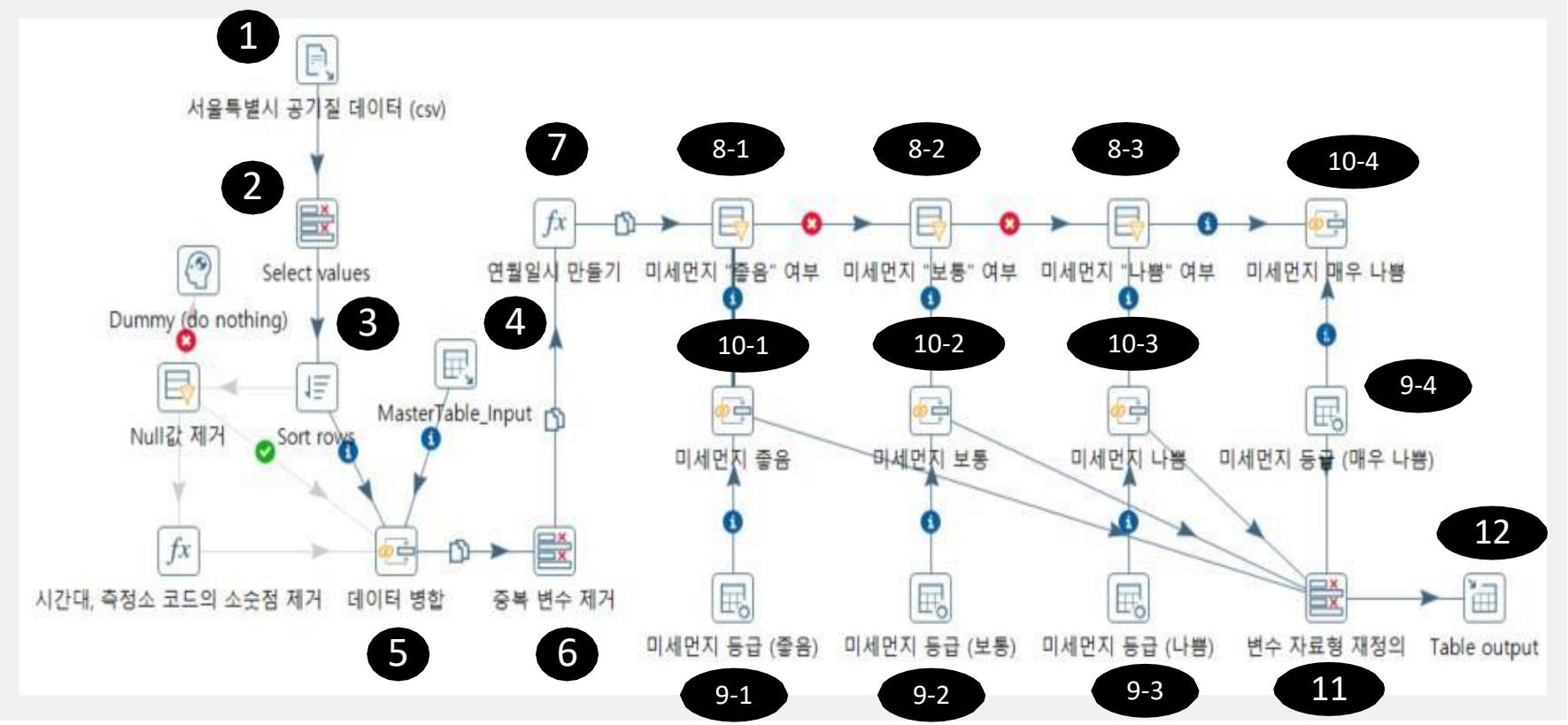
세부 목차

1. 서울시 공기질 데이터 활용
2. 서울시 공기질 데이터 처리

6. 활용예제

6.1 서울시 공기질 데이터 활용

- 서울시 공기질 공공 데이터 활용하여 미세먼지 분석 할 수 있음



6. 활용예제

6.2 서울시 공기질 데이터 처리 및 등급 구분(1/2)

- 서울시 공기질 데이터 분석 프로세스(Step명칭 및 상세 설명)

① Data Input : 공공데이터 포털 서울시 공기질 데이터를 CSV로 다운받은 경로 설정

② Select Value : 필요 변수 선정(PM10,PM20 등 원하는 공기질 지표 선정)

③ Sort rows : 데이터를 Station Code 오름차순으로 정렬

④ Table Input : 좌표 데이터 테이블을 DB에서 가져온다.

⑤ Merge Join : Station Code로 Inner Join (Data Input 과 Table Input)

⑥ Select Value : Merge Join 후 중복 변수 및 불필요 변수 제거 (Address 등)

⑦ Fomula : 시간데이터를 년/월/일/시간으로 분할

⑧ Filter rows : 미세먼지 지수를 필터링(좋음/보통/나쁨/매우나쁨 분류)

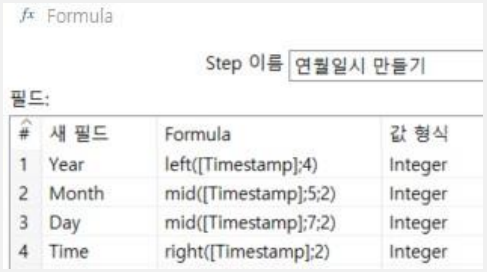
⑨ Add constant rows : 미세먼지 지수 등급 부여(이산화-좋음/보통/나쁨/매우나쁨, 등급으로 표현하기 위함)

⑩ Merge Join : Filter rows 와 Add constant rows Inner Join (미세먼지 지수와 등급을 통합)

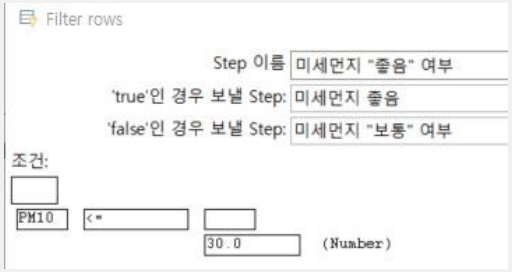
⑪ Select Value : 중복 변수 제거와 메타 변수 정의

⑫ Table output : DB에 테이블 저장

7



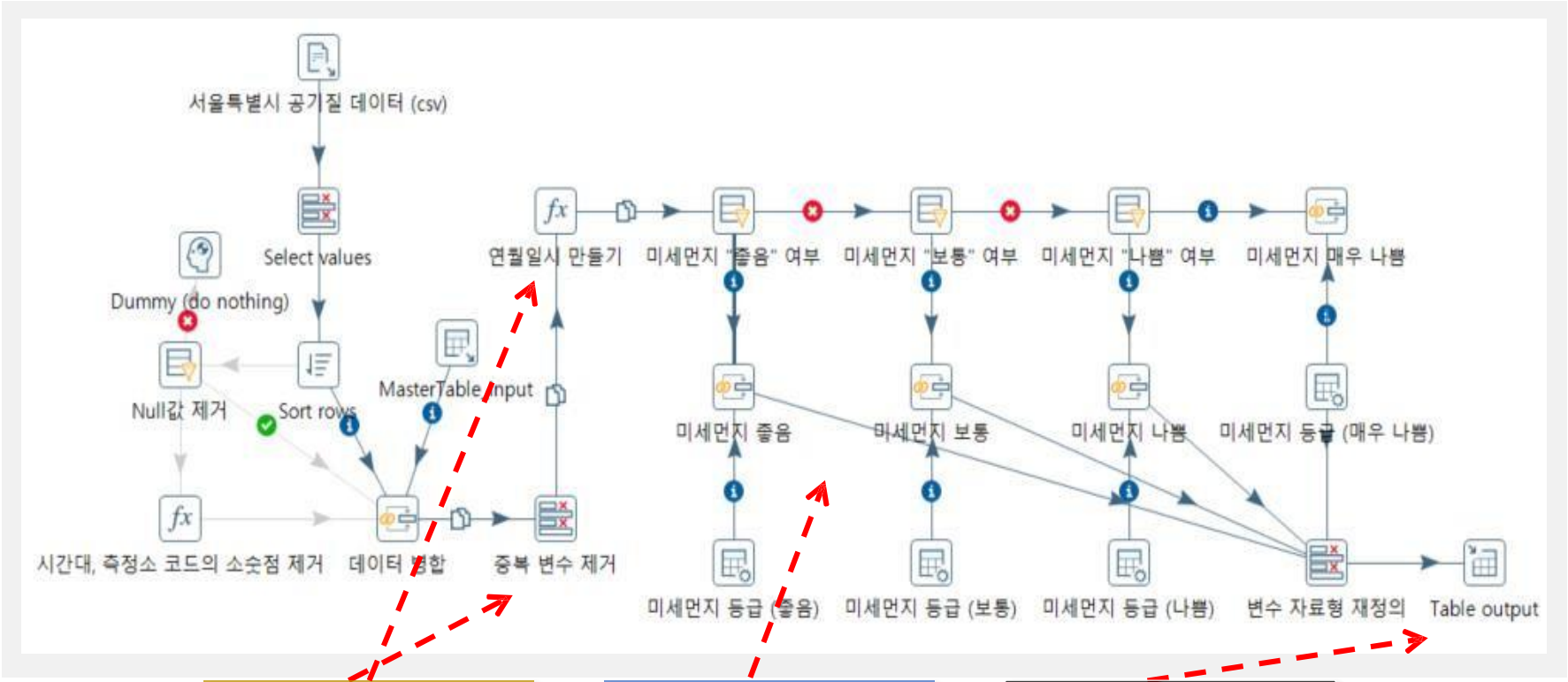
9



6. 활용예제

6.2 서울시 공기질 데이터 처리 및 등급 구분(2/2)

- 필요에 따라 데이터 처리/병합/생성/구분하여 원하는 데이터 분석 결과 도출



7. FAQ

Q **Windows 환경에서 Spoon.bat 시작하면 아무 일도 일어나지 않습니다. 문제 어떻게 분석 할 수 있습니까?**

A Spoon.bat 파일 편집하고 1) 마지막 줄 "start javaw" "java"로 바꾸십시오 2) 다음 줄에 "pause" 추가하십시오 3) 다시 저장하고 다시 시도하십시오. 그런 다음 오류 메시지가 표시되며 다음 질문 통해 이 분석 할 수 있습니다.

Q **매뉴얼에서 행 타입이 섞이지 않을 수도 있다는 것 읽었습니다. 그게 무슨 뜻입니까?**

A 행들이 섞이지 않는다는 것은 단일 홉을 통해 전송되는 모든 행이 동일한 구조, 즉 동일한 필드 이름, 유형, 필드 순서이어야 한다는 것을 의미합니다. 따라서 조건이 행에 대해 true이면 추가 필드를 추가하고, 그렇지 않으면 추가 정보를 추가하려는 경우 (조건에 따라 다른 유형의 행을 얻을 것이므로) 작동하지 않습니다. "안전 모드 사용"을 켜면 런타임에 이를 명시 적으로 확인할 수 있습니다.

8. 용어정리

용어	설 명
Pentaho BI Platform Project	플랫폼 하부구조에서 서비스들의 전달에 초점 맞춘 오픈소스 프로젝트. Pentaho의 end-user 통합과 데이터 통합 기능들 제공. Pentaho BI Platform 프로젝트는 보안, 통합, APIs, 스케쥴링, 워크플로우와 같은 기능들 제공
Pentaho BI Platform	End-user reporting, analysis, back-end 보안 갖는 Dashboard 기능, integration, scheduling, 워크플로우 기능들 지원해 주는 컴포넌트들과 API들 그리고 애플리케이션들 가리킴
Pentaho BI Server	호스트 애플리케이션 서버 내에서 운영하는 J2EE 애플리케이션 이며 사용자 요청에 대한 서비스 제공한다. 이 용어는 어떻게 또는 어디에 배포된다는 것 고려하는게 아니라 플랫폼의 서버 부분 참조하기위해사용됨.
Action Sequence Editor	Action Sequence(Pentaho BI Platform 내에서 동작하는 스크립트) 만들 수 있도록 해주는 이클립스 플러그인
XMI	AXMI Metadata Interchang: XMI는 XML 통해 메타데이터 교환하기 위한 OMG 표준임

Open Source Software Installation & Application Guide