KNME 소개 및 서버 성능평가 자료

본 자료는 '엑셈'사로 부터 받은 자료임을 밝힙니다.

제품 문의는 '엑셈 (http://owleye.co.kr/knime/)'으로 연락 바랍니다.

※ YouTube에 엑셈에서 제작한 강의 있음 ("엑셈 knime" 검색)

Open for Innovation

KNIME

CONTENTS

1

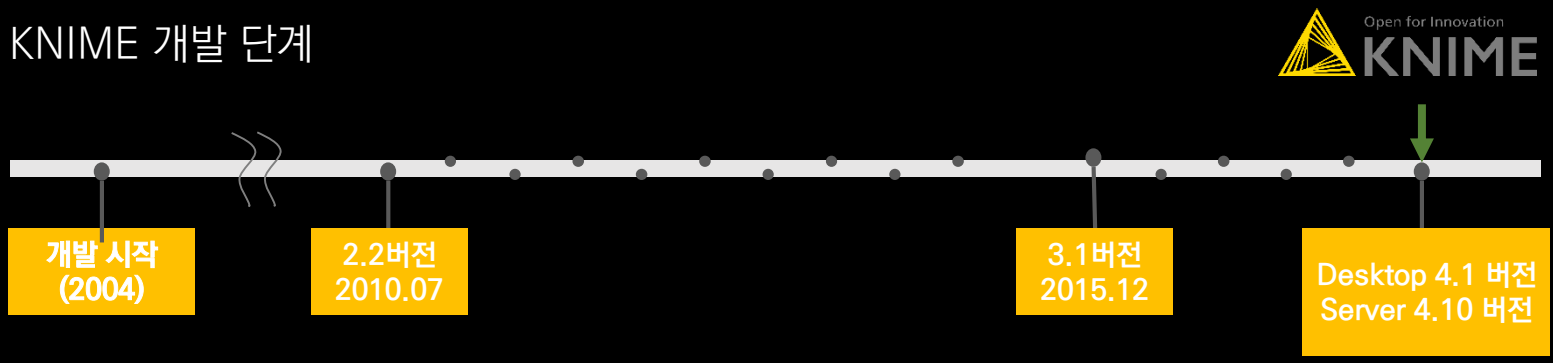
PART ONE

KNIME 소개

KNIME(Konstanz Information Miner)이란

- 워크플로우 기반으로 데이터 선택부터 전처리, 변환, 분석, 평가. 시각화 까지
  일련의 분석 과정을 손쉽게 작업할 수 있는 Total 분석 플랫폼으로
  독일의 Konstanz University에서 소프트웨어 엔지니어 팀이 Java로 개발

- KNIME 개발 단계

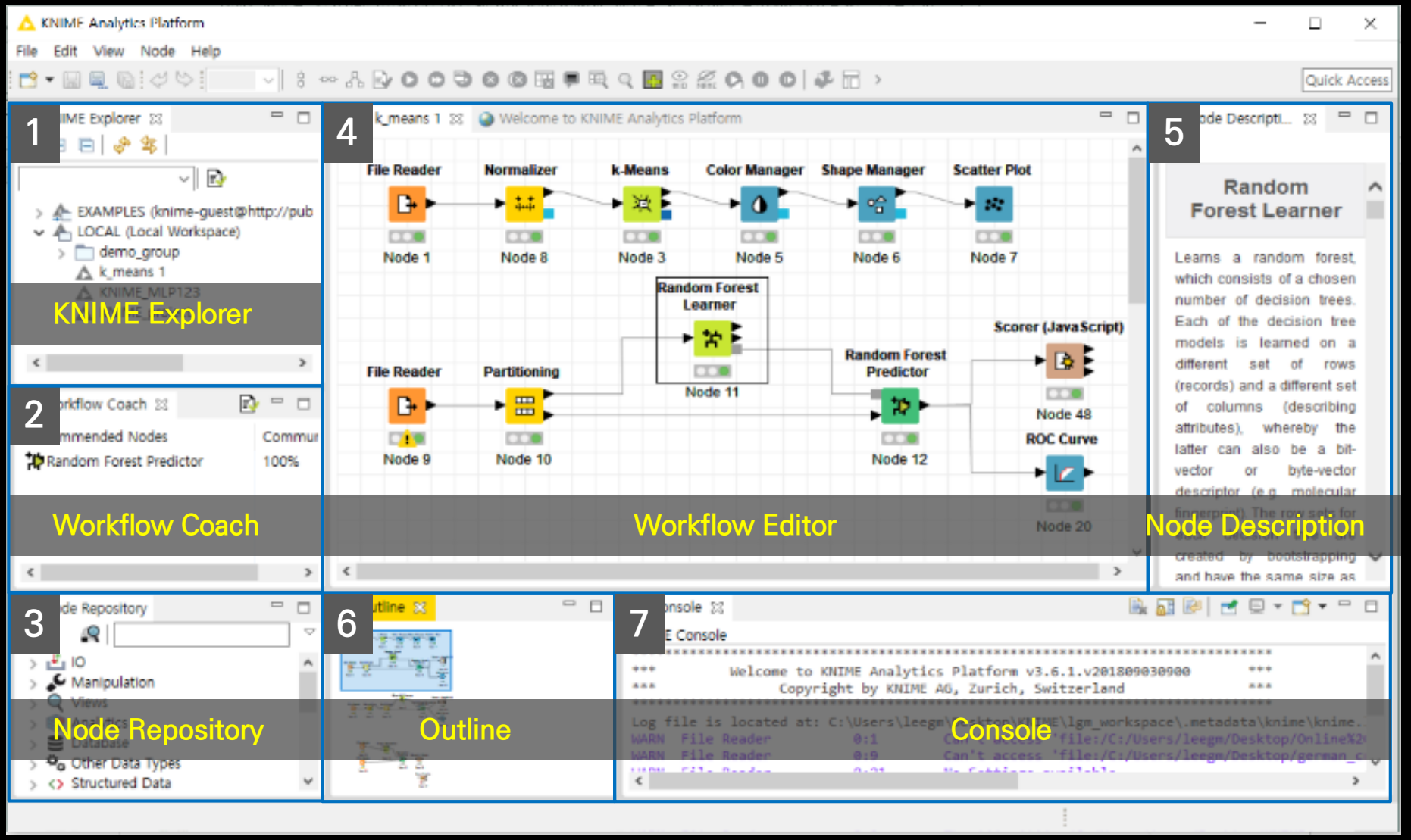


2019 Magic Quadrant

Open for Innovation
KNIME

"데이터 사이언스와 머신러닝 플랫폼" 분야에서 KNIME 이 시장 지속적으로 손 꼽히고 있음

KNIME의 워크플로우 구성 화면

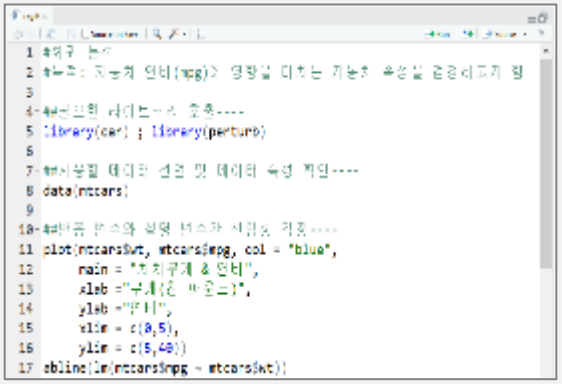


KNIME과 다른 분석 툴과의 차이점

CUI 기반

Character **U**ser **I**nterface
명령어를 작성하여 작업

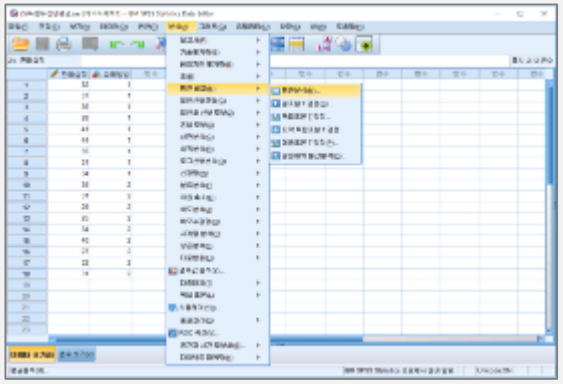
프로그램 : RStudio



GUI 기반

Graphical **U**ser **I**nterface
마우스 클릭으로 작업

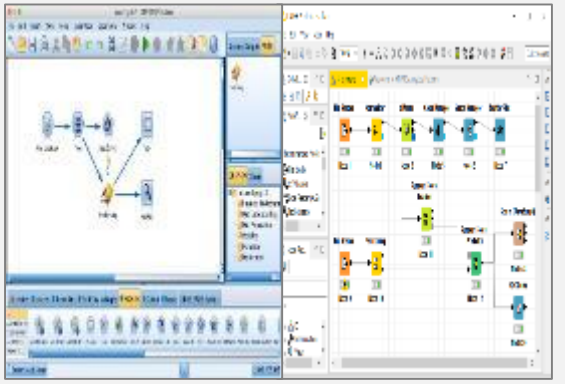
프로그램 : SPSS Statistics



VPL 기반

Visual **P**rogramming **L**anguage
워크플로우를 작성하여 작업

프로그램 : SPSS Modeler, KNIME

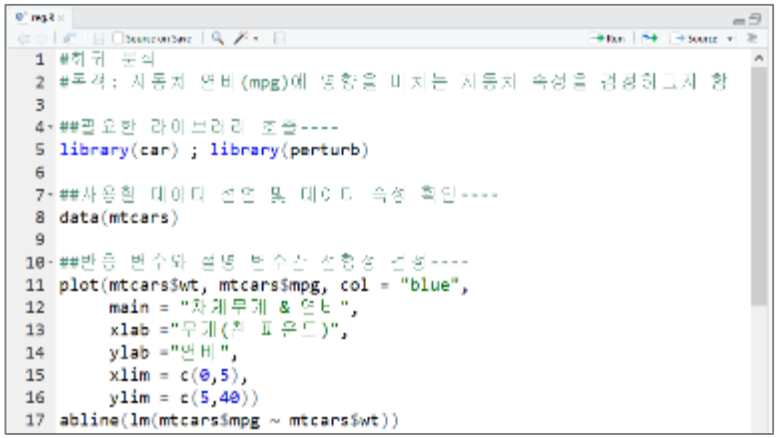


모듈 의존도는 높아지지만, 학습 난이도가 낮으며 분석 과정의 이해가 쉬움

KNIME과 다른 분석 툴과의 차이점

오픈소스 소프트웨어 vs 오픈소스 소프트웨어

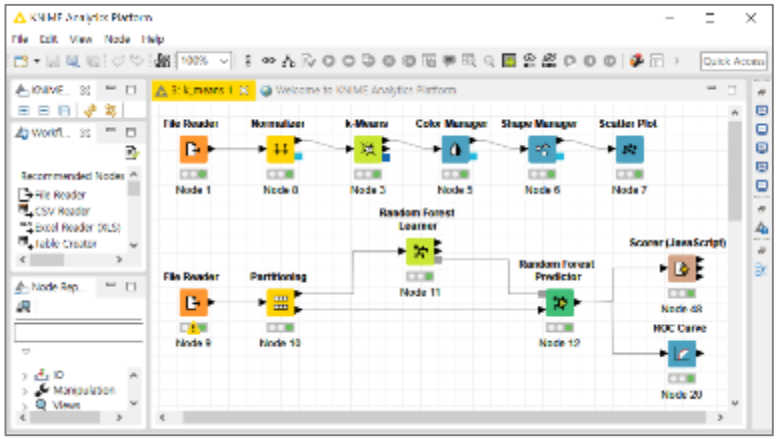
RStudio



```
#회귀 분석
#목적: 자동차 연비(mpg)에 영향을 미치는 자동차 속성을 검정하고자 함

##필요한 라이브러리 로딩----
library(car) ; library(perturb)

##사용할 데이터 선언 및 데이터 속성 확인----
data(mtcars)

##반응 변수와 설명 변수의 산점성 검정----
plot(mtcars$wt, mtcars$mpg, col = "blue",
     main = "차체무게 & 연비",
     xlab ="무게(천 파운드)",
     ylab ="연비",
     xlim = c(0,5),
     ylim = c(5,40))
abline(lm(mtcars$mpg ~ mtcars$wt))
```

- 명령어로 작업이 진행됨
- 분석 진행 단계를 주석을 통해 확인할 수 있음
- 분석기법별 함수는 존재하지만 추가 가능 옵션에 대해서는 옵션을 알고있어야 적용할 수 있음

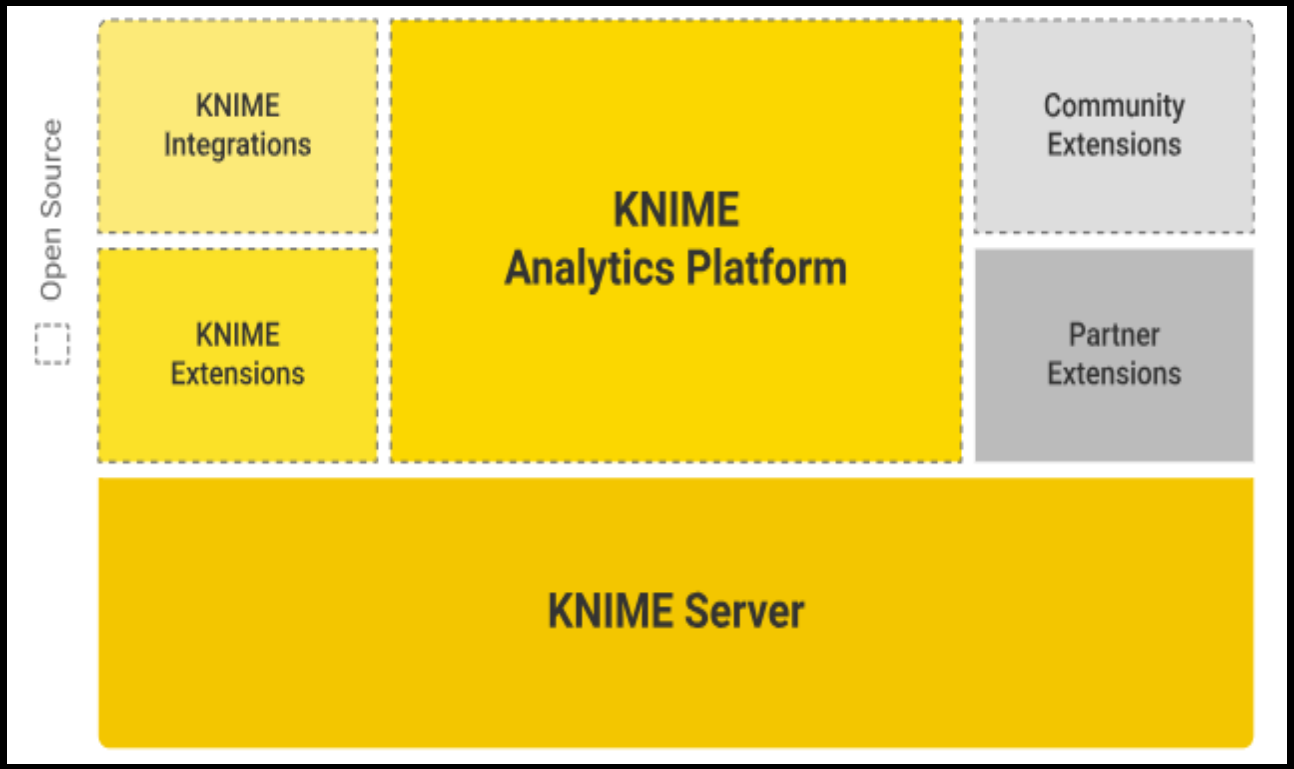
KNIME



- 워크플로우 작성으로 작업이 진행됨
- 분석 진행 단계를 화면에서 바로 확인할 수 있음
- 분석기법별 노드가 존재하며, 분석기법별 노드에 적용할 수 있는 옵션이 확인 가능함

KNIME 제품 구성

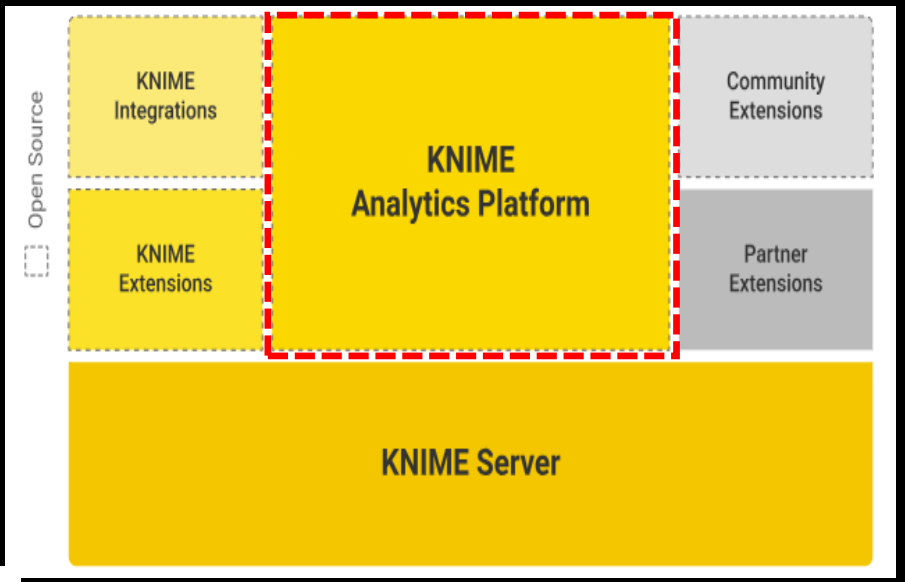
KNIME 6가지로 구성



KNIME 제품 구성

KNIME Analytics Platform

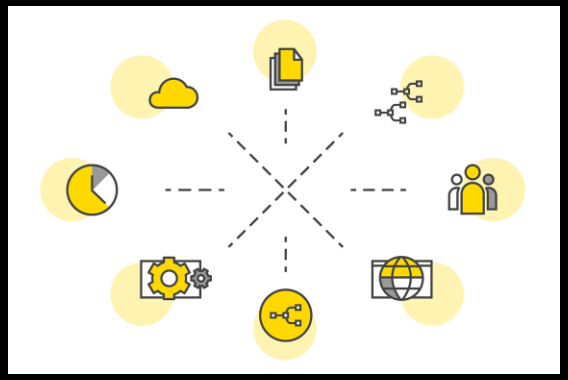
데이터 과학 응용 프로그램 및 서비스를 구현하기 위한
오픈 소스 소프트웨어이며,
지속적으로 새롭게 개발, 통합되어
데이터를 이해하고 활용 할 수 있도록
직관적인 워크플로우를 제공하고
누구나 재사용 가능하도록 구성 요소가 설계되었습니다.

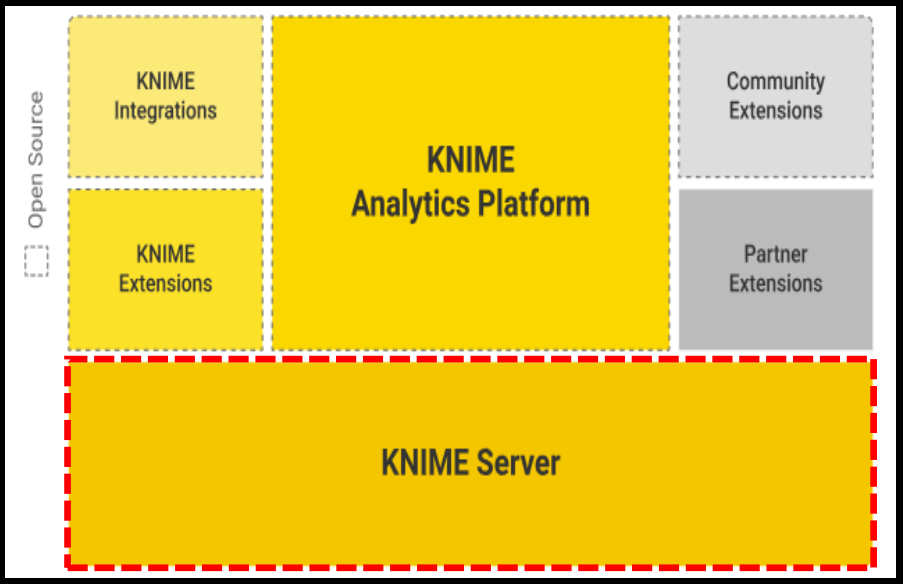


KNIME 제품 구성

KNIME Server

팀 기반 협업, 자동화, 관리, 배포를 위한
워크플로우, 데이터, 분석 가이드를 제공하는
엔터프라이즈 소프트웨어입니다.
비전문가는 KNIME WebPortal을 통해 분석 결과에
접근할 수 있으며, 또한 REST API를 사용하여
분석 응용 프로그램 서비스 및 실시간 IoT 시스템과도
통합 할 수 있습니다.





KNIME 선택의 이유(1/2)

오픈 소스 및 상용 서버 소스 제공

오픈소스 Desktop 소프트웨어로 무료
서버 소프트웨어의 경우 유료

개방적

다양한 데이터(정형데이터, 비정형데이터)에
대해 분석이 가능

직관적

워크플로우 기반으로 분석의 과정을 한눈에 관리

효율성

드래그-앤-드롭(drag-and-drop) 방식으로 분석
초보자도 쉽게 사용

접근성

데이터베이스 테이블, 하둡 파일, 하이브 테이블,
형식있는 파일 등 다양한 소스에 접근이 가능

다양성

약 2,000개의 노드로 다양한 데이터 전처리 및
분석이 가능

언어 사용

R, Python, Java 등과 같은 언어도 사용이 가능

자동화

워크플로우에 대한 스케줄을 등록하여
자동화된 분석이 가능

KNIME 선택의 이유(2/2)

다양한 라이브러리와 소스가 계속 생겨난다.

만들 때 마다 사소한 실수가 많고 쉽게 찾을 수가 없다.

만들어 놓은 모델을 관리하고 배포하기가 쉽지 않다.

협업하기 위한 공유와 스케쥴 관리가 쉽지 않다.

데이터 처리와 모델링 코딩하는 시간이 많이 소요된다.

2 PART THREE

KNIME 성능 평가

KNIME Server

KNIME Server 특징 및 설명

특징 및 설명	Small	Medium	Large
협업			
워크플로우 공유 및 접근 권한 제어	O	O	O
User가 대부분의 일반적인 기능을 재사용할 수 있도록 구성 요소(metanode, workflow...) 등을 업로드 및 공유	O	O	O
메타노드를 암호화하여 컨텐츠 보안 및 지적 재산권 보호			O
Custom Node repository를 사용하여 편의성 및 규정 준수 보장 (TeamSpace)			O
자동화			
서버 또는 클러스터에서 워크플로우 실행	O	O	O
특정 시간에 또는 정기적으로 워크플로우를 실행하거나 또는 리포트를 출력하도록 예약	O	O	O
안전한 환경에서 구성된 하드웨어를 활용할 수 있도록 KNIME 서버에서 워크플로우 수정 및 실행	O	O	O
분산 실행기를 사용하여 다중 머신에서 워크플로우 실행			O
빅데이터 워크플로우를 원격으로 실행하여 KNIME 워크플로우에서 Apache Hadoop 및 Spark에 접근			O
배포			
Consumer에게 KNIME WebPortal 또는 REST API를 통해 분석 애플리케이션 및 서비스에 대한 접근	User만 가능	추가적인 Consumer	무제한의 Consumer
Guided Analytics 생성 및 배포	O	O	O
REST API를 통해 워크플로우를 배포하여 다른 애플리케이션에서 액세스 가능		O	O
관리			
소규모 팀을 위한 User 접근 권한 관리	O	O	O
워크플로우 스냅샷 생성 및 이전 버전과 비교	O	O	O
서버 작업 모니터링(실행 및 예약된 작업), 사용 권한 조정, 진행 중인 서비스 관리	O	O	O
여러 명의 KNIME Analytics Platform에서 서버에 접근하기 위한 중앙 집중식 IT 운영 및 커스터마이징		O	O
회사의 LDAP/Active Directory 설정과 인증 통합 및 사용 권한 관리			O
일반			
서버 설치 수	1	1	여러 대
티켓 시스템 또는 포럼을 통한 전문 지원 서비스 구독	Forum/Email	Email	Email

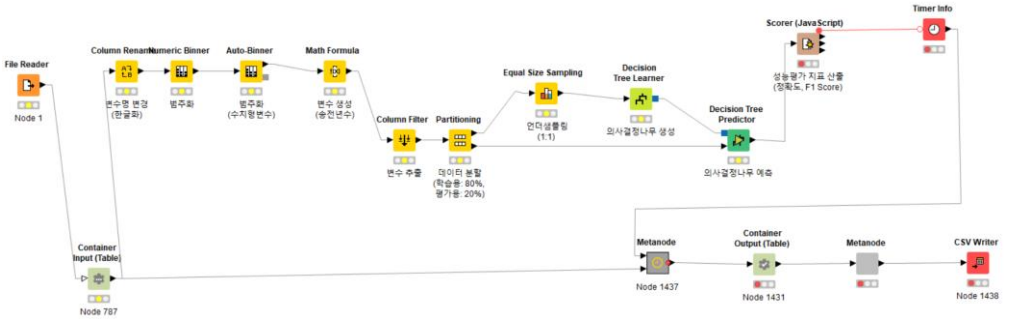
- 스케줄, 워크플로우 기동·공유, 사용자 인증 등 기본 기능 이상의 기능을 제공
  - ✓ Large의 경우 (여러 대 설치 가능) : 빅데이터 Spark 연계 및 LDAP 기반 사용자 계정 관리, WebPortal 및 레포지토리 개인화 가능
  - ✓ Medium/Large의 경우 : Rest API 제공 및 응용 서비스 개발 가능

KNIME Server와 KNIME Analytics Platform 성능 비교

성능 비교를 위한 시뮬레이션 방법

1. 목　적
   · KNIME Server와 KNIME Analytics Platform(AP)에서 대용량 데이터 처리 속도 및 활용 리소스 비교
2. 비교 방법
   · 데이터 크기
      - 145MB (Row : 1,000,000건, 변수 : 22개)
      - 1.42GB (Row : 10,000,000건, 변수 : 22개)

· 대상 Workflow
   - 데이터 처리
   - 의사결정나무 모델 학습 및 평가
   - 데이터 로드 시간은 측정에서 제외



· 시뮬레이션

구분	시뮬레이션 방법
KNIME Server (분산처리)	Job 2개를 동시에 실행, 각각의 서버에 Job 1개씩 할당하여 마지막 Job이 완료된 시간 측정
KNIME Server	Job 2개를 동시에 실행하여 Job이 시작되는 시점부터 마지막 Job이 완료된 시간 측정
KNIME AP	Job 2개를 동시에 실행하여 Job이 시작되는 시점부터 마지막 Job이 완료된 시간 측정

KNIME Server와 KNIME Analytics Platform 성능 비교

시뮬레이션 결과

구분			KNIME AP	KNIME Server	KNIME Server (분산 처리)	
					Server 1	Server 2
OS			Windows10 (64bit)	Ubuntu 16.04 LTS (64bit)	Ubuntu 16.04 LTS (64bit)	Ubuntu 16.04 LTS (64bit)
조건	활용 스레드 수		4	4	4	4
	최대 설정 메모리 (GB)		8	8	8	8
결과	평균 사용 메모리 (GB)		3.9	3.4	1.2	0.9
	실행 시간 (분)	평균	4.1	1.8	1.6	
		최소	3.7	1.7	1.5	
		최대	4.9	1.8	1.6	

* 데이터 크기 : 145MB (1,000,000건)

- 100 만 건 데이터 기준 KNIME Server가 KNIME AP 보다 처리 시간이 평균 2.5 배 빠른 것으로 나타남
- 100 만 건 데이터 기준 KNIME Server(분산 처리)가 KNIME Server 보다 처리 시간이 빠른 것으로 나타남

KNIME Server와 KNIME Analytics Platform 성능 비교

시뮬레이션 결과

구분			KNIME AP	KNIME Server	KNIME Server (분산 처리)	
					Server 1	Server 2
OS			Windows10 (64bit)	Ubuntu 16.04 LTS (64bit)	Ubuntu 16.04 LTS (64bit)	Ubuntu 16.04 LTS (64bit)
조건	활용 스레드 수		4	4	4	4
	최대 설정 메모리 (GB)		32	32	32	32
결과	평균 사용 메모리 (GB)		측정 불가	4	8	2
	실행 시간 (분)	평균		21.2	19.5	
		최소		21.0	19.2	
		최대		21.4	20.0	

* 데이터 크기 : 1.42GB (10,000,000건)

- 1,000 만 건 데이터 기준 KNIME AP 에서는 워크플로우를 실행할 수 없어 처리 시간 측정 불가능
- 1,000 만 건 데이터 기준 KNIME Server(분산 처리)가 KNIME Server 보다 빠른 것으로 나타남

OWEYE

KNIME DataRobot Contents News Contact 로그인 Enter your search

Contact

owleye@ex-em.com

Any questions?
Send us a message

Your name

Your e-mail

Subject

Message

Send a message

Contact Us

제품 구매 문의 관련사항은 owleye@ex-em.com 로 연락바랍니다.

Q&A

Open for Innovation

KNIME