04

비정형데이터분석

주요 프로그래밍 언어의 이해

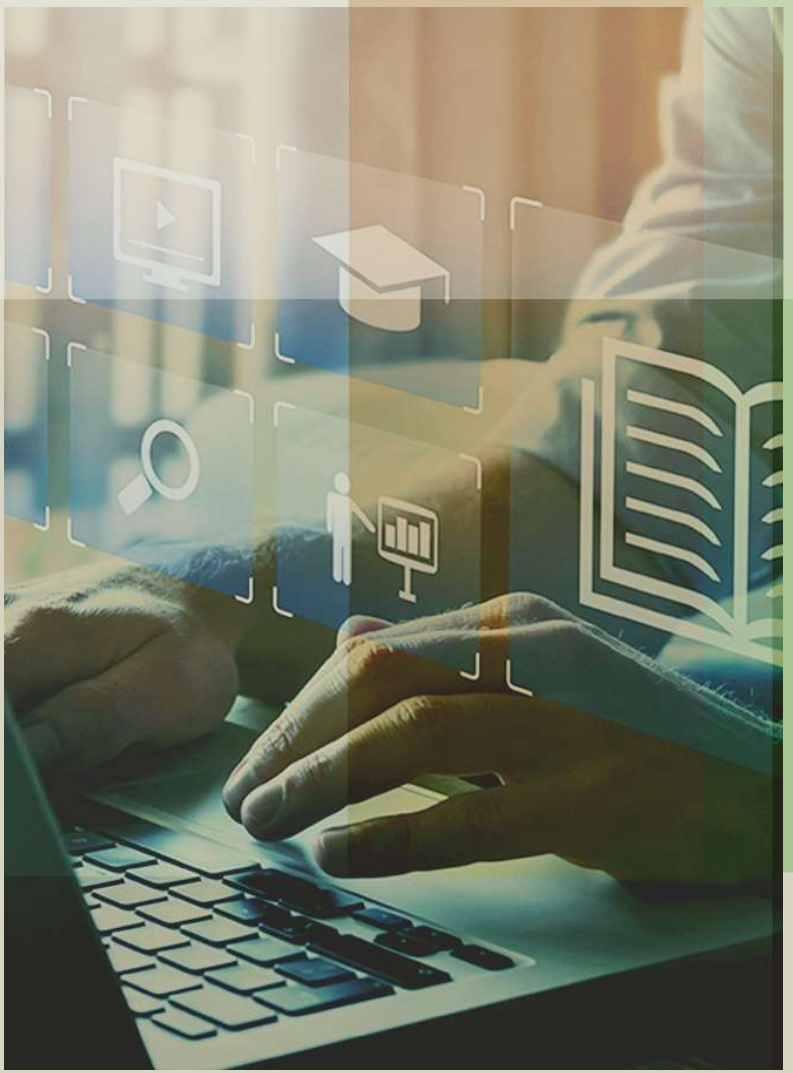
통계·데이터과학과 장영재 교수

학습목차

1. 주요 프로그래밍 언어의 이해(2)
2. 프로그래밍 언어의 선택

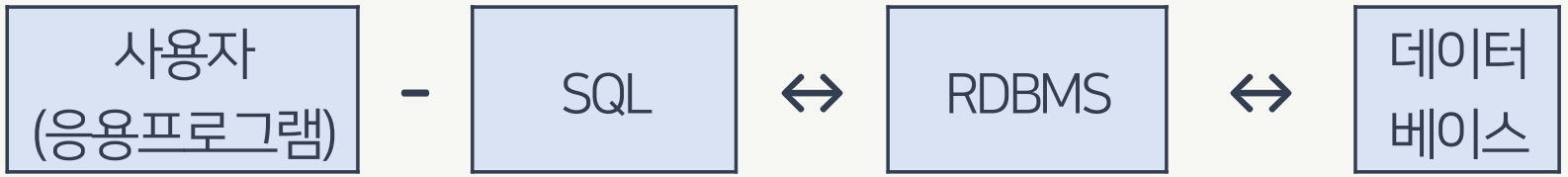
01

주요 프로그래밍 언어의 이해(2)



1. 주요 프로그래밍 언어의 이해 - ① SQL의 개요

- SQL(Structured Query Language)은 관계형 데이터베이스 관리시스템(Relational Database Management System, DBMS)을 관리하는 목적으로 제안된 특수 목적 프로그래밍 언어



<그림> SQL과 관계형 데이터베이스 관리 시스템(RDBMS)

- SQL은 데이터베이스 관리 시스템에 관한 언어 기반의 인터페이스
- SQL은 사용방식에 따라 일반적으로 대화형과 내장형으로 구분
  - → 대화형은 각 행 별로 명령어가 입력되는 방식. 즉, 사용자가 SQL 명령어를 직접 입력할 때, 데이터베이스 관리 시스템이 이를 해석하고 처리하여 결과를 반환
  - → 내장형은 타 프로그래밍 언어(호스트 언어) 내에서 SQL 명령어가 사용되는 방식

1. 주요 프로그래밍 언어의 이해 - ① SQL의 개요

- SQL의 세 가지 구성요소는 데이터 정의어, 데이터 조작어, 데이터 제어어
  - SQL 데이터 정의어(DDL) ex) CREATE, DROP, ALTER
    - → 데이터베이스에 저장될 데이터에 대한 형식, 구조, 제약조건들을 명시
  - SQL 데이터 조작어(DML) ex) SELECT, INSERT, DELETE, UPDATE
    - 특정 데이터 검색 질의, 데이터베이스의 갱신, 삽입, 삭제 등을 관리
  - 데이터 제어어(DCL) ex) GRANT, REVOKE, COMMIT, ROLLBACK
    - 데이터베이스 접근, 갱신, 삽입, 삭제 등 작업이 정확하게 수행되어 무결성 유지

1. 주요 프로그래밍 언어의 이해 - ① SQL의 개요

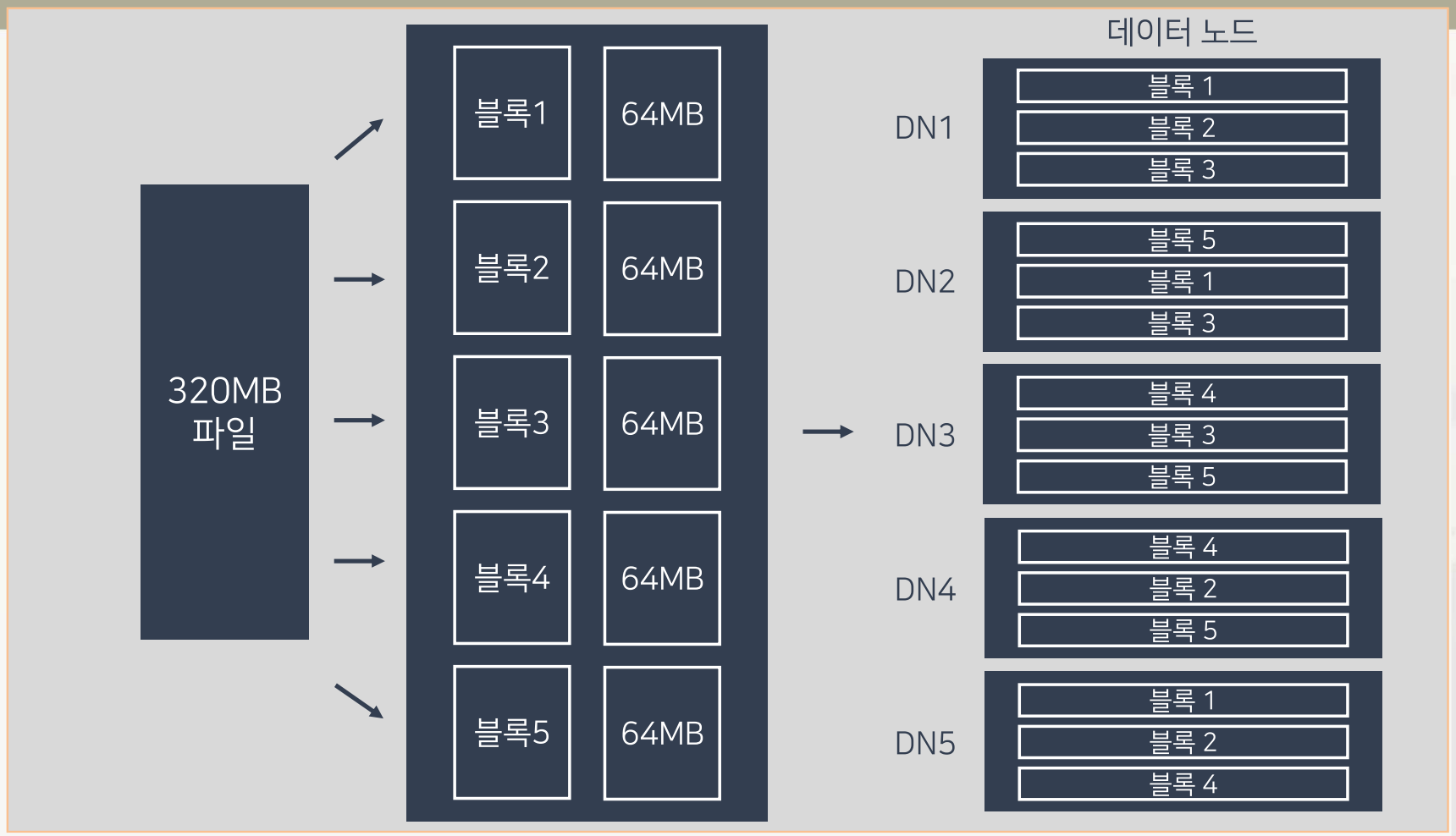
- SQL 문법은 모든 종류의 데이터베이스 연산에 응용이 가능하며 정수, 실수, 문자 등 여러 가지 자료형태를 지원하는 특징이 있음
  - 과거에는 데이터베이스가 바뀌면 새로운 언어를 배워서 관리를 해야 했지만, SQL의 등장은 이러한 제약을 해소

1. 주요 프로그래밍 언어의 이해 - ② Hadoop의 개요

- Hadoop은 야후(Yahoo)의 Doug Cutting에 의해 개발된 대규모 데이터를 처리, 분석할 수 있는 Java 기반의 오픈소스 프레임 워크
  - Hadoop은 아파치(Apache) 재단의 오픈소스로 공개 - Apache Hadoop
- 분산 파일 시스템(Hadoop Distributed File System;HDFS)과 맵리듀스(MapReduce)가 Hadoop을 특징짓는 두 가지 요소

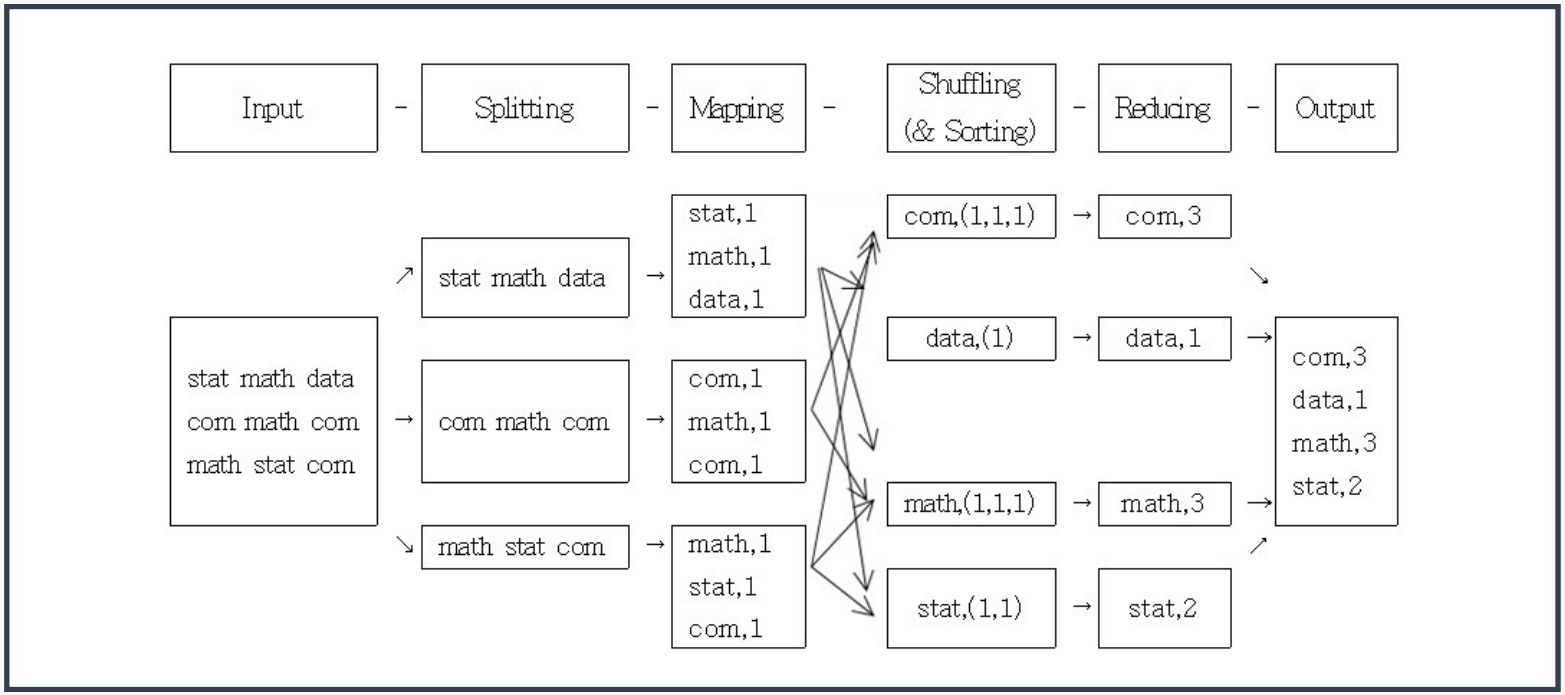
1. 주요 프로그래밍 언어의 이해 - ② Hadoop의 개요

- 분산 파일 시스템(Hadoop Distributed File System;HDFS)
  - → HDFS은 대용량의 파일을 분산된 서버에 저장하고 그 데이터를 빠르게 처리할 수 있게 설계
  - → 데이터의 안정성과 무결성(입력이나 변경을 막고 읽기만 허용하여 일관성을 유지) 등이 특징
  - → HDFS는 데이터 저장 시 복제 데이터를 저장하여 데이터 유실을 방지하고 분산 서버 간 오류를 주기적으로 검증
  - → 클라이언트가 끊임없이 데이터에 접근하는 스트리밍 방식으로 많은 양의 데이터를 처리



<그림> HDFS에서 데이터가 저장되는 방식

- 맵리듀스(MapReduce)는 대량의 텍스트 데이터를 받아들여 문자열 데이터를 분리하고 정리하여 정보를 추출하는 과정



<그림> 맵리듀스의 사례

1. 주요 프로그래밍 언어의 이해 - ③ Java의 개요

- Java는 Sun Microsystem사에서 개발한 객체지향 프로그래밍 언어로서 가전제품에 탑재할 프로그램 개발 목적으로 개발된 Oak라는 언어를 모태로 함
  - 2010년 Oracle 사에서 Sun을 인수하면서 Java의 개발과 관리, 배포 등을 담당
    - → 웹 애플리케이션 개발에 가장 많이 사용되는 언어 중 하나로 손꼽히며 모바일 기기용 소프트웨어 개발을 위해서도 많이 사용

1. 주요 프로그래밍 언어의 이해 - ③ Java의 개요

- Java는 운영체제나 하드웨어 의존하지 않는 특징
  - 하드웨어에 종속되지 않는 중간 파일*인 바이트코드(Bytecode)에 기인
    - 프로그래밍 언어와 기계어의 중간 정도에 위치하기 때문에 붙여진 명칭
  - Java의 독립성은 JVM* (Java Virtual Machine)과 같은 가상 머신을 거쳐 실행되는 방식 때문에 유지될 수 있음
    - JVM을 설치하려면 통상 JDK(Java software Development Kit)라고 불리는 개발 도구를 설치

      → JVM을 이용하면 스마트폰, 컴퓨터와 같은 서로 다른 하드웨어나 Windows, Linux 등 서로 다른 운영체제에서 동일한 프로그램을 개발할 수 있음

Java 프로그램	=	Java 프로그램	=	Java 프로그램
JVM1	≠	JVM2	≠	JVM3
Mac OS	≠	Windows	≠	Linux
하드웨어1	≠	하드웨어2	≠	하드웨어3

<그림> JVM의 특징

Java class file(프로그램명.java) ↓ public class 이름과 동일 Java compiler (Javac) ↓ Java 바이트 코드 (프로그램명.class)			컴파일 과정
JVM for Mac (인터프리터 /JIT compiler) ↓ Mac OS	≠ ≠	JVM for Winodws (인터프리터 /JIT compiler) ↓ Windows	실행과정

<그림> JVM의 작동 방식

1. 주요 프로그래밍 언어의 이해 - ③ Java의 개요

- Java는 확장성이 있고, 응용 및 효율적 실행이 가능한 특징이 있음
  - Java에는 머신 러닝 및 데이터 과학을 위한 수많은 라이브러리와 툴이 있음
  - 확장성이 좋아 응용 프로그램을 확장할 때 많이 사용
  - 인터넷 또는 네트워크를 통하여 효율적으로 실행 가능한 분산 환경에 적합

1. 주요 프로그래밍 언어의 이해 - ④ SAS의 개요

- SAS는 1966년 미국 노스캐롤라이나 주립대학교에서 통계분석에 특화된 통계패키지의 형태로 개발(유료 프로그램)
  - 보고서 작성, 프로그래밍 등 기법에 경영진단 기법, 툴 박스의 추가 등으로 그 사용 목적과 범위가 넓어지면서 종합정보처리 시스템의 성격으로 변모
  - 큰 대용량의 데이터를 읽기 쉬우며 데이터 구조 변경이 용이
  - 분석 결과의 요약이나 리포팅 등 보고서 작성까지 가능
  - 그래픽 기능이 뛰어나며 엑셀, 오라클 등의 다양한 포맷의 데이터를 읽어 활용 가능

1. 주요 프로그래밍 언어의 이해 - ④ SAS의 개요

- SAS/CORE, SAS/BASE와 SAS/STAT 및 SAS/GRAPH 등 단위 소프트웨어로 구성
  - 기본 단위 소프트웨어 외에도 추가 기능 구현이 가능한 다양한 소프트웨어가 존재

SAS/AF 소프트웨어 : 상호작용이 원활한 GUI 환경 구축

SAS/ETS : 시계열 자료의 분석, SAS/IML : 행렬 연산 , SAS/QR : 품질관리

SAS/INSIGHT 및 Enterprise miner : 데이터마이닝, SAS/PH는 : clinical trial

1. 주요 프로그래밍 언어의 이해 - ④ SAS의 개요

- SAS의 프로그램(코드)은 크게 Data step과 Proc step으로 구성
  - Data step에서 생성된 데이터는 라이브러리(library)라는 곳에 저장 (임시 또는 영구)
  - 편집기 창(Editor)에서 모든 명령문들을 입력하여 작성한 코드가 실행되면 로그(log) 창에서 코드 오류 검증 결과, 출력창(Output)에 실행 결과가 나타남
- 2014년 SAS 사는 교육용 목적으로 SAS University Edition을 무료로 배포하기 시작
  - 기본적인 통계 분석과 데이터 시각화, 결과물 출력 등이 가능

02

프로그래밍 언어의 선택

2. 프로그래밍 언어의 선택

- 많이 사용되는 프로그래밍 언어 중 분석 목적에 맞는 적절한 언어를 선택
  - 비정형 데이터 중에서 텍스트 데이터가 차지하는 비중이 매우 크므로 분석의 범주를 텍스트 데이터에 맞추어 선택
  - 텍스트 데이터 분석을 위해서는 분석이 용이하고 쉽게 접근할 수 있는 파이썬과 R이 주로 활용됨
    → 설치 환경과 실습 여건 등을 고려하여 본 교재에서는 R을 선택하여 활용

다음시간안내

05

텍스트 데이터 불러오기