제2장 비정형 데이터 분석의 도구

1. 비정형 데이터의 분석

일반적인 데이터 분석 과정에서도 **데이터의 수집이나 전처리 과정이 중요**

비정형 데이터 분석의 경우에는 훨씬 더 그 중요성이 부각

2. 분석 도구의 구현

2.1 프로그래밍의 의의

비정형 데이터의 분석 도구를 개발하고 사용함에 있어서 가장 강조되는 능력

→ 프로그래밍 능력

2.2 프로그래밍 언어 선택의 기준

프로그래밍 언어를 선택하는 기준

① 프로그래밍 목적

② 사용 환경

③ 성능

2.3 빅데이터 시대의 프로그래밍 언어

R, Python, SQL

3. 주요 프로그래밍 언어의 이해

3.1 파이썬의 개요

인터프리터 / 오픈소스 / 객체지향 / 동적 타이핑

3.2 R의 개요

파이썬과 유사 / 통계분석에 특화

3.3 SQL의 개요

데이터베이스 관리시스템을 제어할 목적으로 만들어진 특수 목적 프로그래밍 언어

- 데이터 정의어(DDL): CREATE, DROP, ALTER
- 데이터 조작어(DML): SELECT, INSERT, DELETE, UPDATE
- 데이터 제어어(DCL): GRANT, REVOKE, COMMIT, ROLLBACK

3.4 Hadoop의 개요

대규모 데이터를 처리, 분석할 수 있는 오픈소스 프레임 워크(Java 기반)

하둡 분산 파일 시스템(Hadoop Distributed File System; HDFS)과 맵리듀스(MapReduce)가 Hadoop을 특징짓는 두 가지 요소

3.5 Java의 개요

썬마이크로시스템즈에서 개발한 객체지향 프로그래밍 언어

중간파일인 바이트코드를 만들고, 그것을 자바가상머신(JVM; Java Virtual Machine)으로 실행하기 때문에 운영체제나 하드웨어 종류에 관계없이 동작

3.6 SAS의 개요

유료

4. 프로그래밍 언어의 선택

Python 또는 R 중에 하나 선택 + SQL