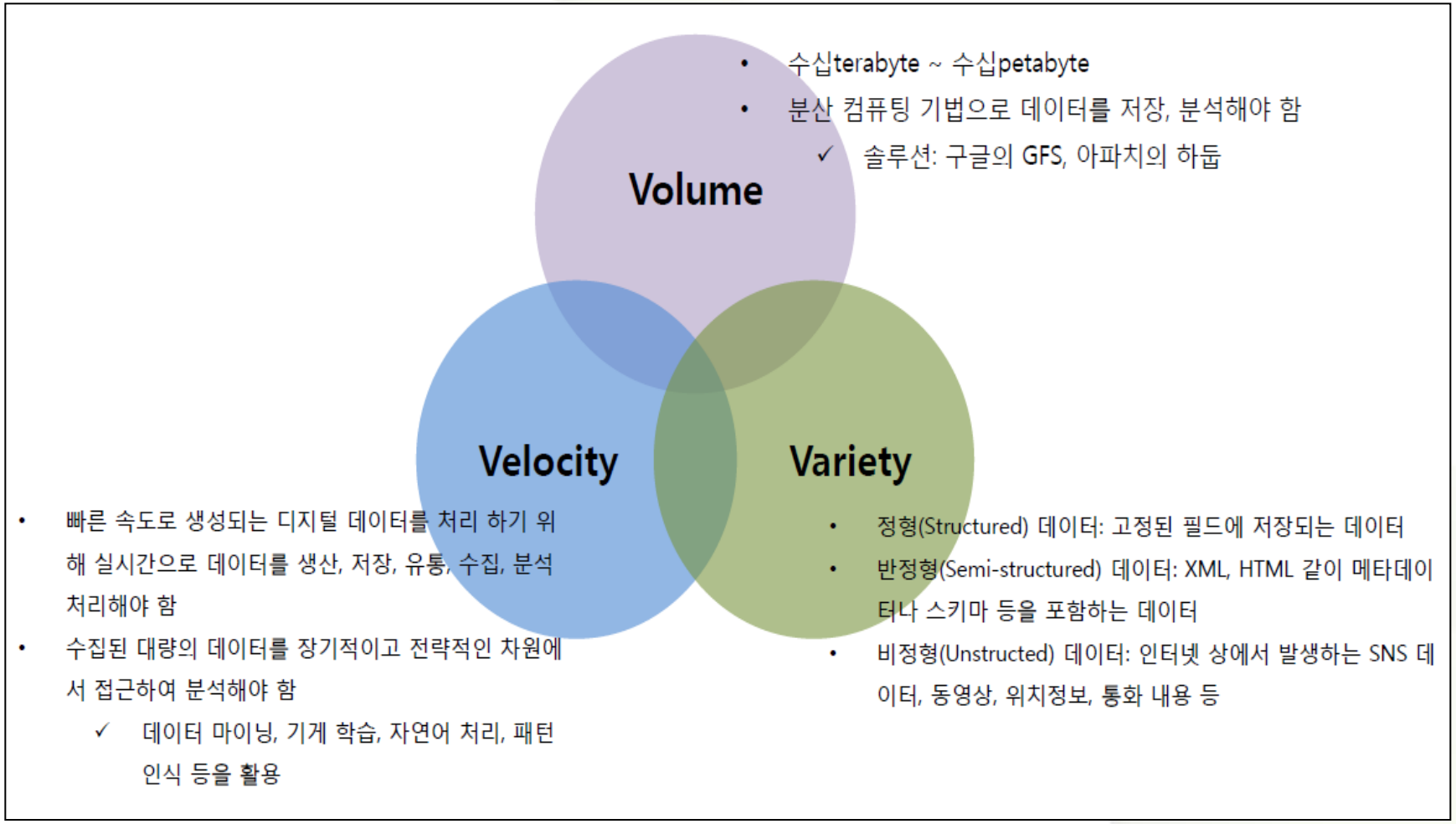
Introduction to Spark

빅데이터

- 빅데이터의 3대 요소
  - 크기 (Volume), 속도 (Velocity), 다양성 (Variety)

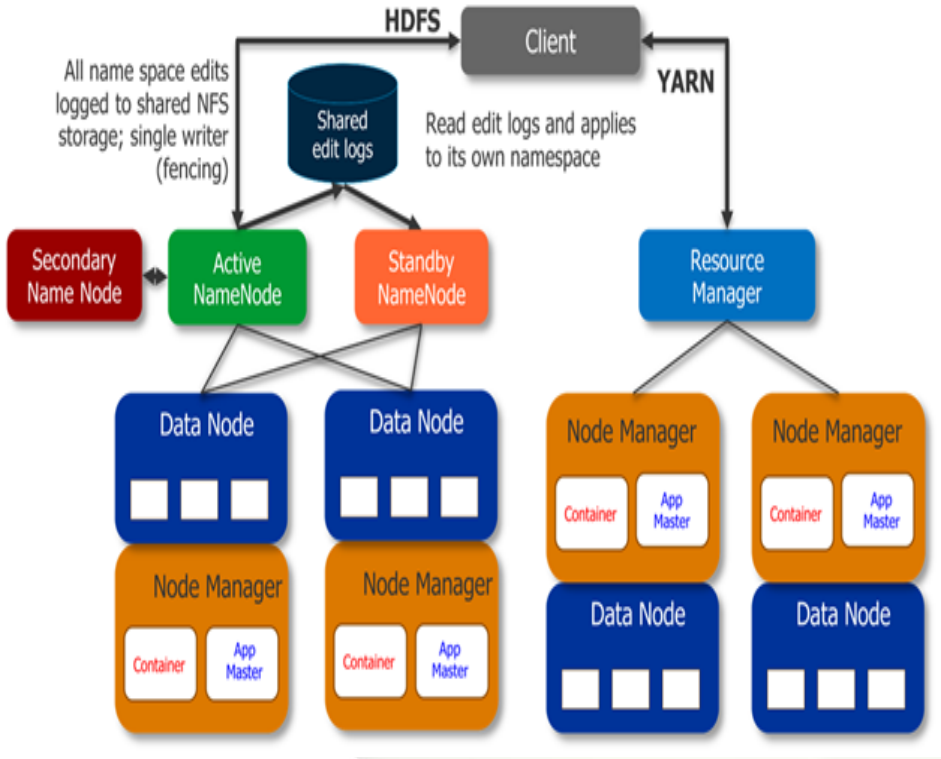


하둡 (Hadoop)

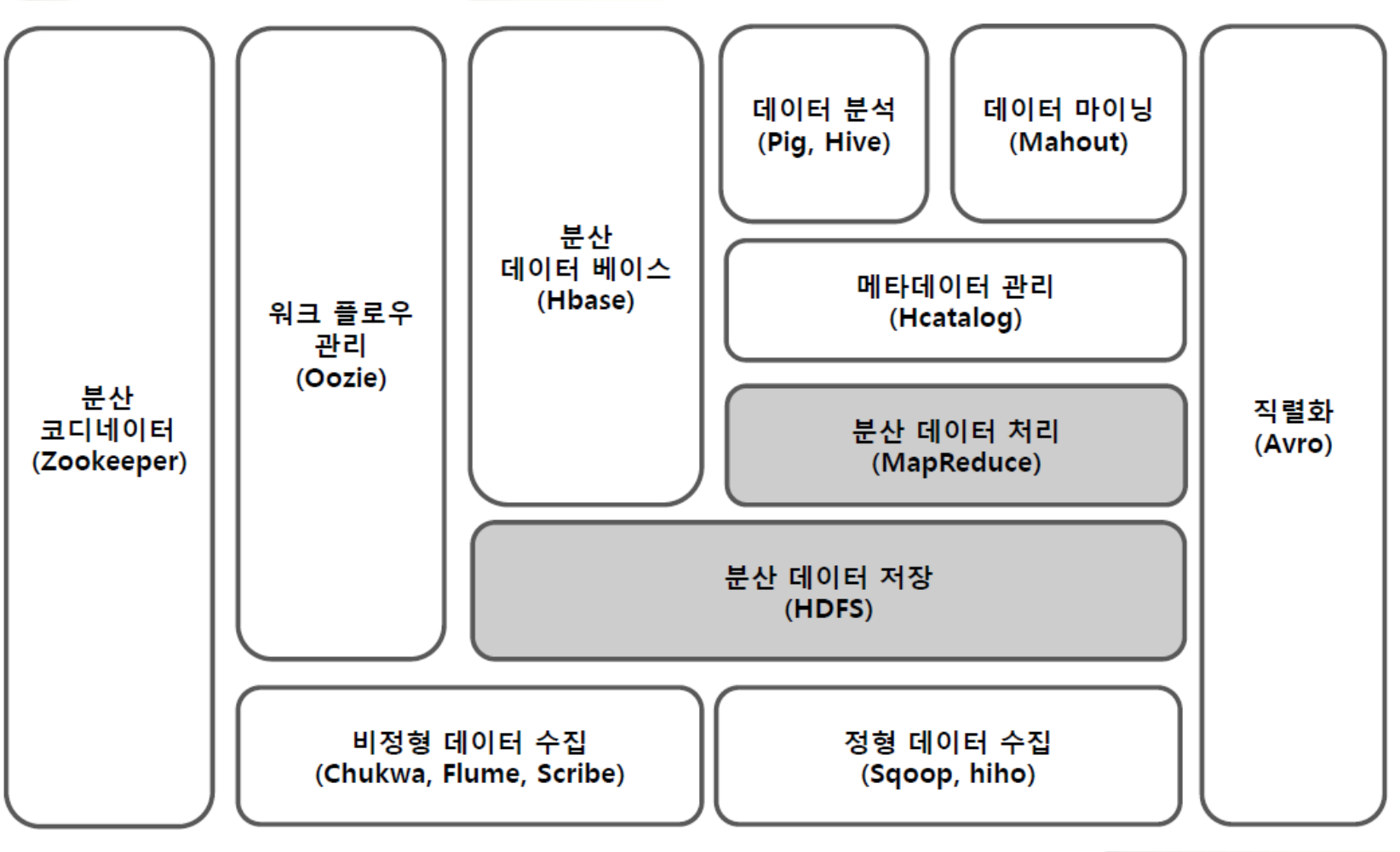
- 빅데이터를 분산 처리할 수 있는 자바 기반의 오픈소스 프레임워크

- 분산 파일 시스템인 HDFS(Hadoop Distributed Files System)에 데이터를 저장하고 분산 처리 시스템인 맵리듀스를 이용해 데이터를 처리

- 2005년에 더그 커팅(Doug Cutting)이 구글이 논문으로 발표한 GFS(Google File System)와 MapReduce를 구현한 결과물



하둡 에코시스템


분산
코디네이터
(Zookeeper)
워크 플로우
관리
(Oozie)
분산
데이터 베이스
(Hbase)
데이터 분석
(Pig, Hive)
데이터 마이닝
(Mahout)
메타데이터 관리
(Hcatalog)
분산 데이터 처리
(MapReduce)
분산 데이터 저장
(HDFS)
직렬화
(Avro)
비정형 데이터 수집
(Chukwa, Flume, Scribe)
정형 데이터 수집
(Sqoop, hiho)

하둡의 단점

- 대부분의 연산 작업을 파일시스템 기반으로 처리 → 상대적으로 낮은 성능
- 복잡한 데이터 분석 요구사항을 맵과 리듀스 패턴만으로 해결하기 어려움
- 자바 언어 기반으로 파이썬, R 등 다른 분석용 도구와 연동이 어려움
- SQL on Hadoop 계열의 도구와 같이 맵리듀스를 편리하게 구현할 수 있는 도구들이 있지만 데이터 분석 요구사항을 충분히 반영하는데 한계가 있음

스파크 (Spark)

- 하둡 기반 맵리듀스의 단점을 보완하기 위해 개발된 분산 데이터 처리 환경
- 메모리를 이용한 데이터 저장 방식을 제공함으로써 머신러닝 등 반복적인 데이터 처리가 필요한 분야에서 높은 성능 구현
- 최적화 과정을 통해 효율적인 데이터 처리 및 성능 향상 가능
- 자연스럽고 강력한 다수의 데이터 처리 함수 제공 → 프로그램의 복잡도를 현저하게 낮춤
- 자바, 스칼라, 파이썬, R을 사용해서 스파크 애플리케이션 개발 가능
- Spark SQL, MLlib 등 다양한 데이터 처리 분야에 특화된 라이브러리 제공

스파크 데이터 모델

- RDD (Resilient Distributed Dataset)
  - 스파크 내부에 존재하는 분산 데이터에 대한 모델
  - 다수의 서버에 분산 방식으로 저장된 데이터 요소들의 집합
  - 병렬 처리 및 장애 복구 가능

- DataFrame(after 1.3), DataSet(after 1.6)
  - Column들로 구성된 Schema를 사용하는 데이터 모델
  - 관계형 데이터베이스의 테이블과 유사한 방식의 데이터 처리 모델 제공
  - 향상된 최적화 도구 사용
  - 스파크 2.0부터 DataFrame 클래스는 DataSet 클래스로 통합
    - » R, 파이썬은 DataFrame 형식 사용
    - » 자바는 DataSet 형식 사용
    - » 스칼라는 DataFrame와 DataSet 모두 사용

스파크 실행 환경 구축

- 스파크 설치

교재 27 ~ 33페이지 및 설치 command list 파일 참고

스파크 실행 환경 구축

- 스파크 예제 실행

교재 33 페이지 참고

스파크 실행 환경 구축

- spark-shell 사용

교재 41 페이지 참고

스파크 개발 환경 구축

- JDK 설치(윈도우 7 기준)
  - 다운로드 → https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

Java SE Development Kit 8u221

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

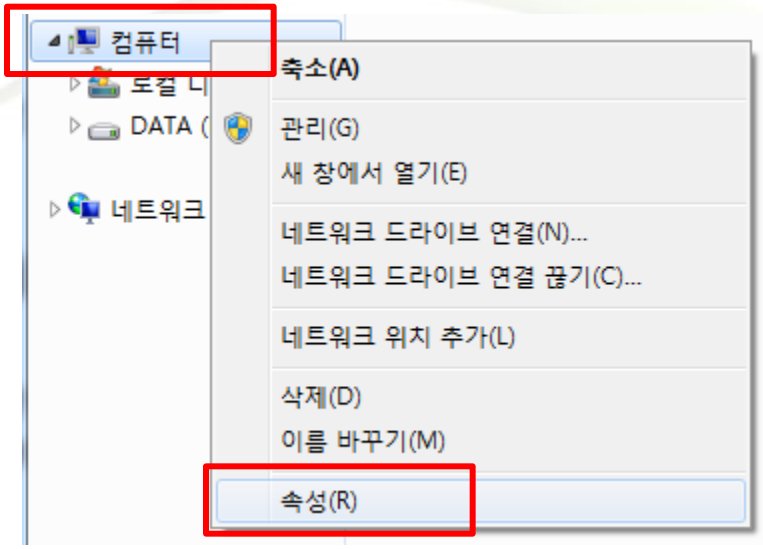
○ Accept License Agreement ◉ Decline License Agreement

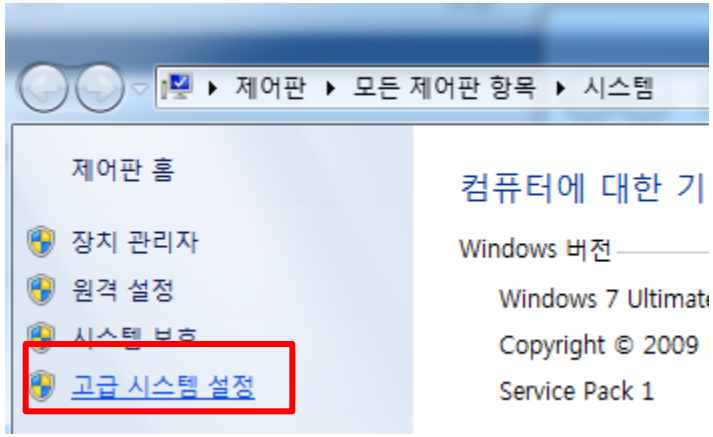
Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u221-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.81 MB	jdk-8u221-linux-arm64-vfp-hflt.tar.gz
Linux x86	174.18 MB	jdk-8u221-linux-i586.rpm
Linux x86	189.03 MB	jdk-8u221-linux-i586.tar.gz
Linux x64	171.19 MB	jdk-8u221-linux-x64.rpm
Linux x64	186.06 MB	jdk-8u221-linux-x64.tar.gz
Mac OS X x64	252.52 MB	jdk-8u221-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.99 MB	jdk-8u221-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.23 MB	jdk-8u221-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.66 MB	jdk-8u221-solaris-x64.tar.Z
Solaris x64	91.95 MB	jdk-8u221-solaris-x64.tar.gz
Windows x86	202.73 MB	jdk-8u221-windows-i586.exe
Windows x64	215.35 MB	jdk-8u221-windows-x64.exe

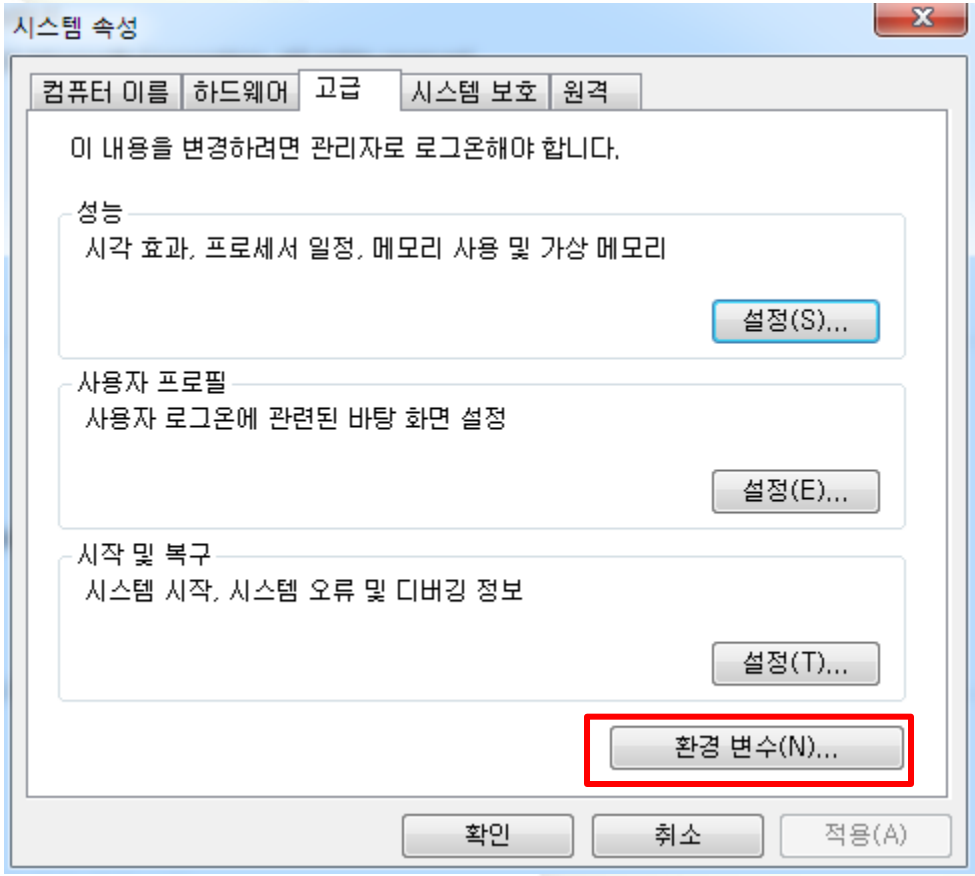
  - 다운로드 후 설치 파일을 관리자 권한으로 실행해서 설치

스파크 개발 환경 구축

- JDK 설치 - 환경 변수 설정 (윈도우 7 기준)

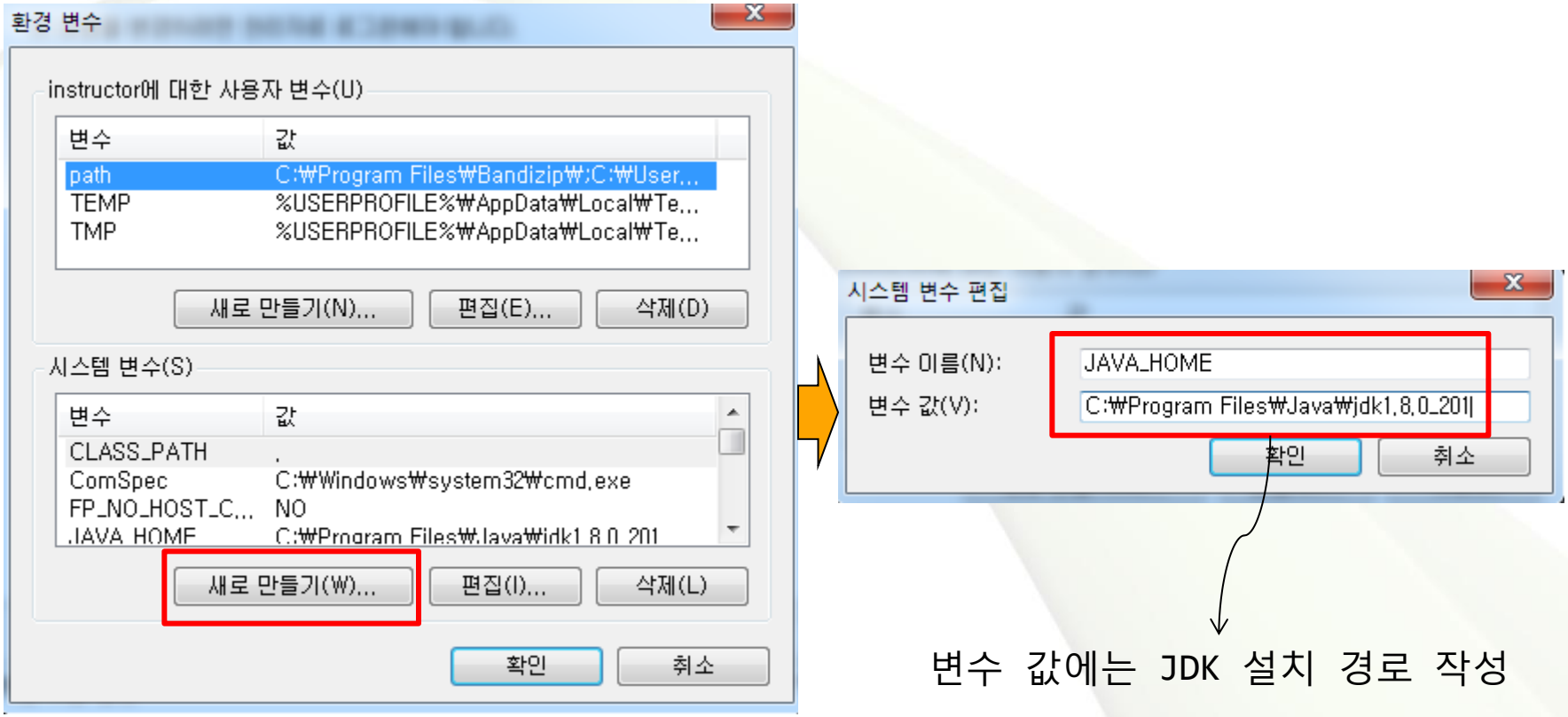






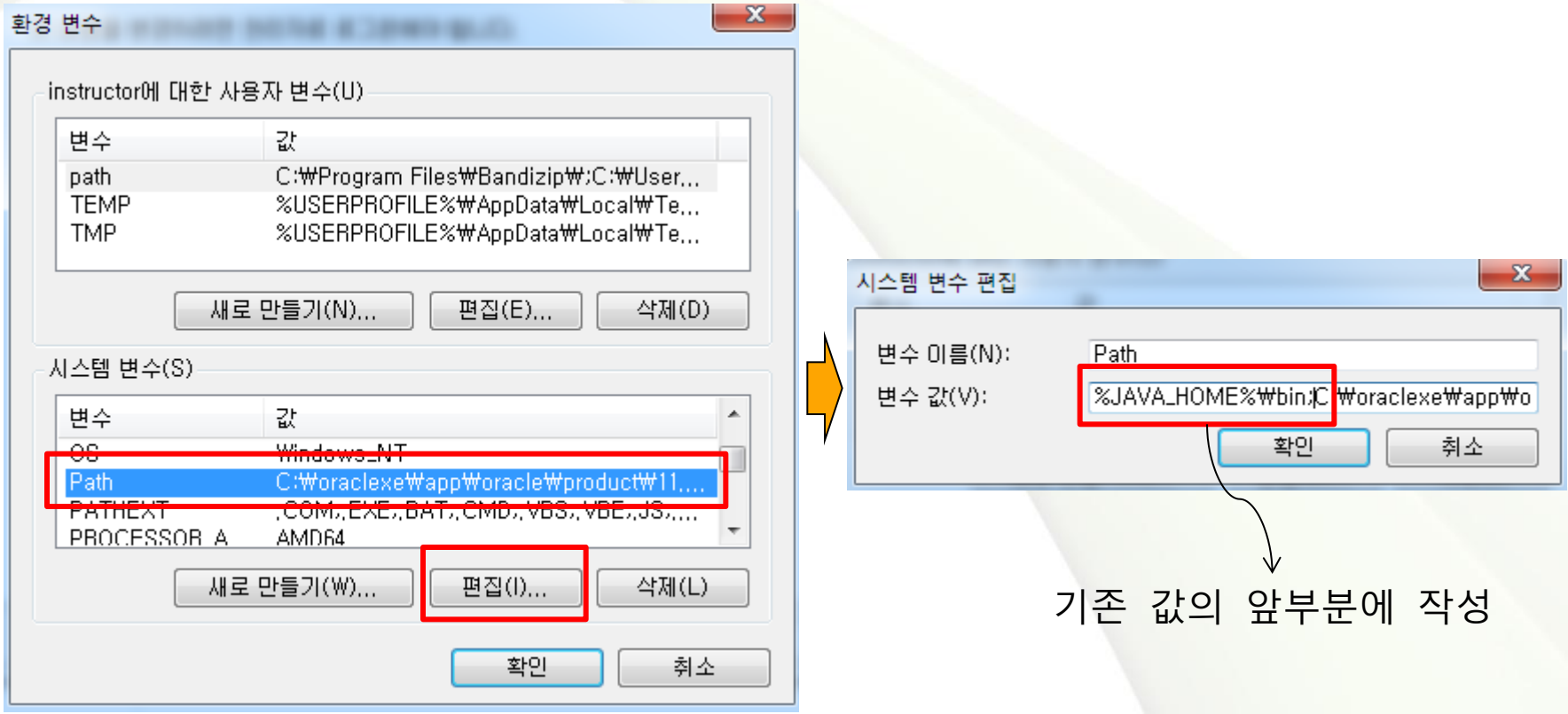
스파크 개발 환경 구축

- JDK 설치 - 환경 변수 설정 (윈도우 7 기준)



스파크 개발 환경 구축

- JDK 설치 - 환경 변수 설정 (윈도우 7 기준)



스파크 개발 환경 구축

- Maven 설치
  - 생략 - 설치하지 않음

- 스파크 설치
  - 실행 환경 구축할 때와 같은 압축 파일 다운로드 및 압축해제
  - 다운로드 → http://mirror.navercorp.com/apache/spark/spark-2.4.3/spark-2.4.3-bin-hadoop2.7.tgz
  - 스파크 설치 경로를 SPARK_HOME 환경 변수로 등록 (JDK 설치 참고)

- 파이썬 설치 (별도 python 설치 문서 참고)
  - Anaconda 또는 Miniconda 설치
  - 이름을 pysparkdev로 지정해서 가상 환경 만들기

스파크 개발 환경 구축

- 스칼라 설치 (생략 - 설치하지 않음)

- ScalaIDE 설치 및 설정

교재 57 ~ 61 페이지 참고

스파크 개발 환경 구축

- 예제 프로젝트 설정

교재 61 ~63 페이지 참고

- 예제 프로젝트 빌드 및 실행

교재 73 ~ 85 페이지 참고

스파크 실행 환경 구축 2

- 파이썬 설치 및 설정 (브라우저의 노트북 환경에서 스파크 사용 목적)
  - Anaconda or Miniconda 설치 (여기서는 Miniconda 설치)
  - 가상 파이썬 환경 만들기
  - 기본 패키지 설치
  - 스파크용 스칼라, SQL 커널 설치
  - PySpark 사용을 위한 환경 변수 설정

command-list 항목3 ~ 5 참고

스파크 실행 환경 구축 2

- 예제 실습

notebook 환경에서 교재 81, 85 페이지 예제 실습