

빅데이터 분석

빅데이터 분석 소개

1. 빅데이터 분석 개요

빅데이터 분석에 대한 기본 정의와 활용 범위를 설명한다.



2. 빅데이터 분석에 활용하는 기술

빅데이터 분석에서 사용할 4가지 기술(임팔라, 제플린, 머하웃, 스쿱)을 소개하고 각 기술별 주요 기능과 아키텍처, 활용 방안을 알아본다.



3. 분석 파일럿 실행 1단계 - 분석 아키텍처

스마트카의 빅데이터 분석과 관련한 요구사항을 구체화하고, 분석 요구사항을 해결하기 위한 파일럿 아키텍처를 제시한다.



4. 분석 파일럿 실행 2단계 - 분석 환경 구성

스마트카의 빅데이터 분석 아키텍처를 실제로 설치 및 환경을 구성한다. 임팔라, 스쿱, 제플린, 머하웃 순으로 설치하게 된다.



5. 분석 파일럿 실행 3단계 - 임팔라를 이용한 데이터 실시간 분석

임팔라로 스마트카 데이터셋을 인메모리 기반으로 실시간 조회 및 분석한다. 또한 하이브에서 사용했던 쿼리를 임팔라에서 실행 및 비교한다.



6. 분석 파일럿 실행 4단계 - 제플린을 이용한 실시간 분석

제플린의 웹 유저 인터페이스를 이용해 스마트카 운행 지역 분석을 위한 스파크 SQL을 작성 및 실행한다.



7. 분석 파일럿 실행 5단계 - 머하웃을 이용한 데이터 마이닝

스마트카 데이터셋을 이용해 머하웃의 3가지 데이터 마이닝을 실행한다. 추천(Recommendation), 분류(Classification), 군집(Clustering) 분석을 진행한다.

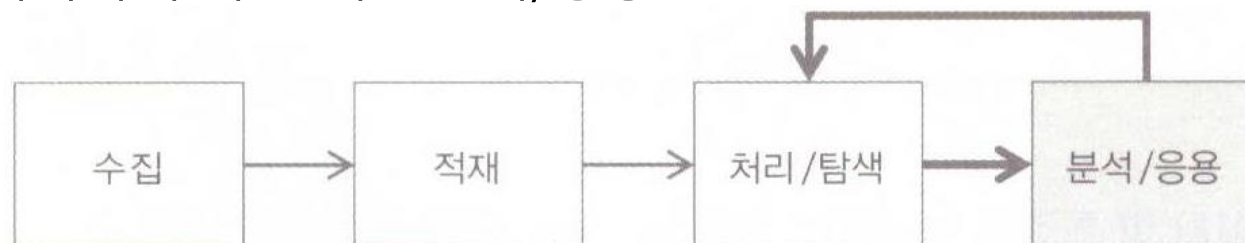


8. 분석 파일럿 실행 6단계 - 스쿱을 이용한 분석 결과 외부 제공

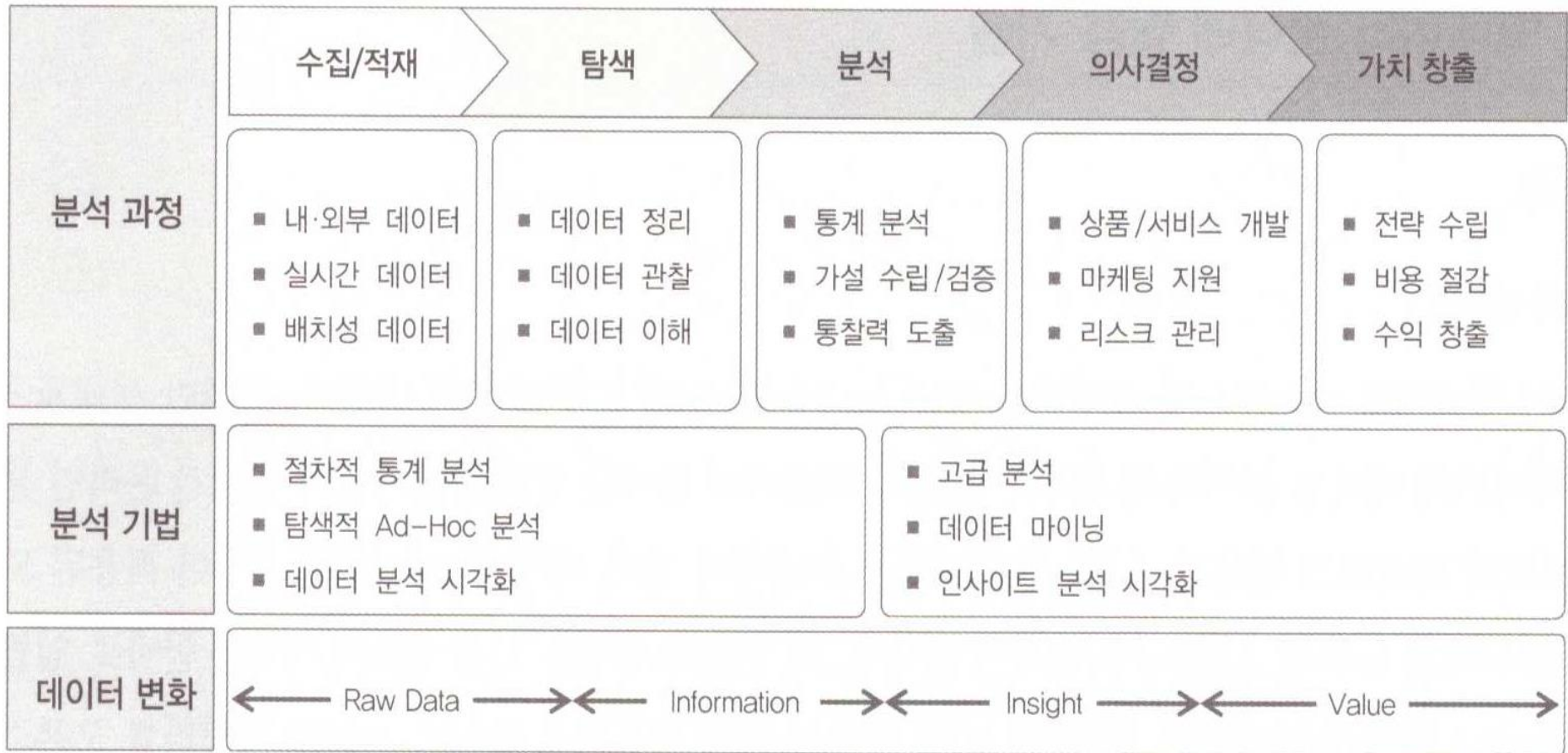
탐색/분석된 스마트카 데이터셋을 외부 RDBMS로 제공한다.

빅데이터 분석 개요

- 탐색과 분석을 반복하며 의미 있는 데이터를 추출해 문제를 명확히 정의하고 해결하는 과정.
- 분석의 목적에 따른 유형
 - 기술 분석: 분석 초기 데이터의 특징을 파악하기 위해 선택, 집계, 요약 등 양적 기술 분석을 수행
 - 탐색 분석: 업무 도메인 지식을 기반으로 대규모 데이터셋의 상관관계나 연관성을 파악
 - 추론 분석: 전통적인 통계분석 기법으로 문제에 대한 가설을 세우고 샘플링을 통해 가설을 검증
 - 인과 분석: 문제 해결을 위한 원인과 결과 변수를 도출하고 변수의 영향도를 분석
 - 예측 분석: 대규모 과거 데이터를 학습해 예측 모델을 만들고, 최근의 데이터로 미래를 예측
- 빅데이터 구축 단계 - 분석/응용



빅데이터 분석 프로세스



빅데이터 분석 활용 기술 - 임팔라(Impala)

➤ 임팔라 소개

- 빅데이터 분석을 인메모리 기반의 실시간 온라인 분석으로까지 확대 기대.
- 구글의 드레멜(Dremel) 논문 2010년에 발표.
- 2012년 10월 실시간 빅데이터 분석 질의가 가능한 임팔라를 클라우데라가 오픈 소스로 발표.

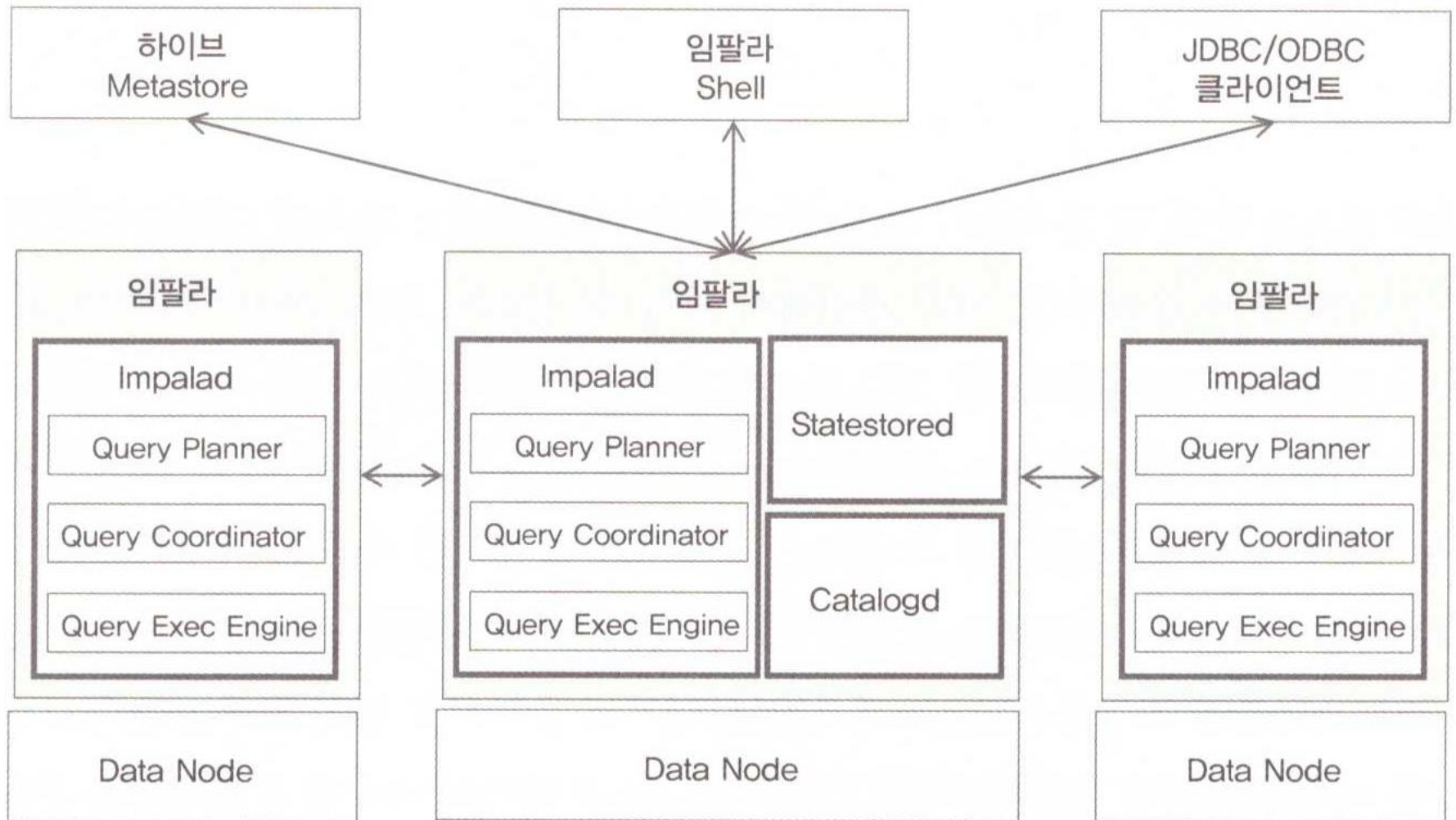
빅데이터 분석 활용 기술 - 임팔라(Impala)

➤ 임팔라 기본 요소

공식 홈페이지		http://impala.apache.org/
주요 구성 요소	Impalad	하둡의 데이터노드에 설치되어 임팔라의 실행 쿼리에 대한 계획, 스케줄링, 엔진을 관리하는 코어 영역
	Query Planner	임팔라 쿼리에 대한 실행 계획을 수립
	Query Coordinator	임팔라 잡리스트 및 스케줄링을 관리
	Query Exec Engine	임팔라 쿼리를 최적화해서 실행하고, 쿼리 결과를 제공
	Statestored	분산 환경에 설치돼 있는 Impalad의 설정 정보 및 서비스를 관리
	Catalogd	임팔라에서 실행된 작업 이력들을 관리하며, 필요 시 작업 이력을 제공
라이선스	Apache	
유사 프로젝트	Tez, Spark SQL, Drill, Tajo	

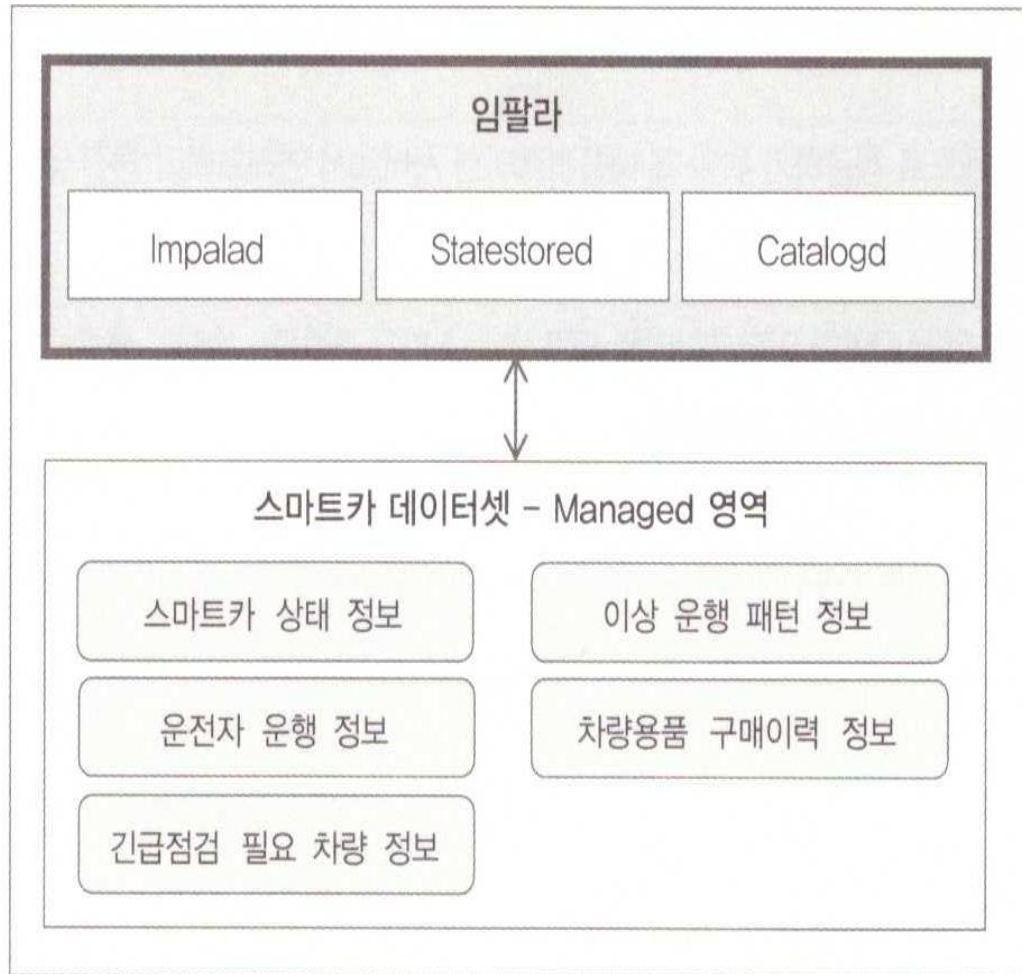
빅데이터 분석 활용 기술 - 임팔라(Impala)

➤ 임팔라 아키텍처



빅데이터 분석 활용 기술 - 임팔라(Impala)

➤ 임팔라 활용 방안



- Impala Engine을 이용, 대용량 데이터를 실시간으로 Ad-Hoc 분석
- Managed 영역의 스마트카 데이터셋을 빠르게 탐색

빅데이터 분석 활용 기술 - 제플린(Zeppelin)

➤ 제플린 소개

- 대용량 데이터를 효과적으로 탐색 및 분석하기 위해서는 대용량 데이터셋을 빠르게 파악하고 이해하기 위한 분석 및 시각화 툴 필요.
- 하둡의 저장소에 있는 데이터를 참조하여 데이터 분석이 가능하도록 스파크를 기반으로 하는 제플린이 탄생.
- 국내 스타트업 기업인 NFLaps에서 2013년부터 주도하고 있는 오픈소스 프로젝트로, 2014년 12월 아파치 재단에 인큐베이팅됐고, 2016년 5월 아파치 최상위 프로젝트로 승격.

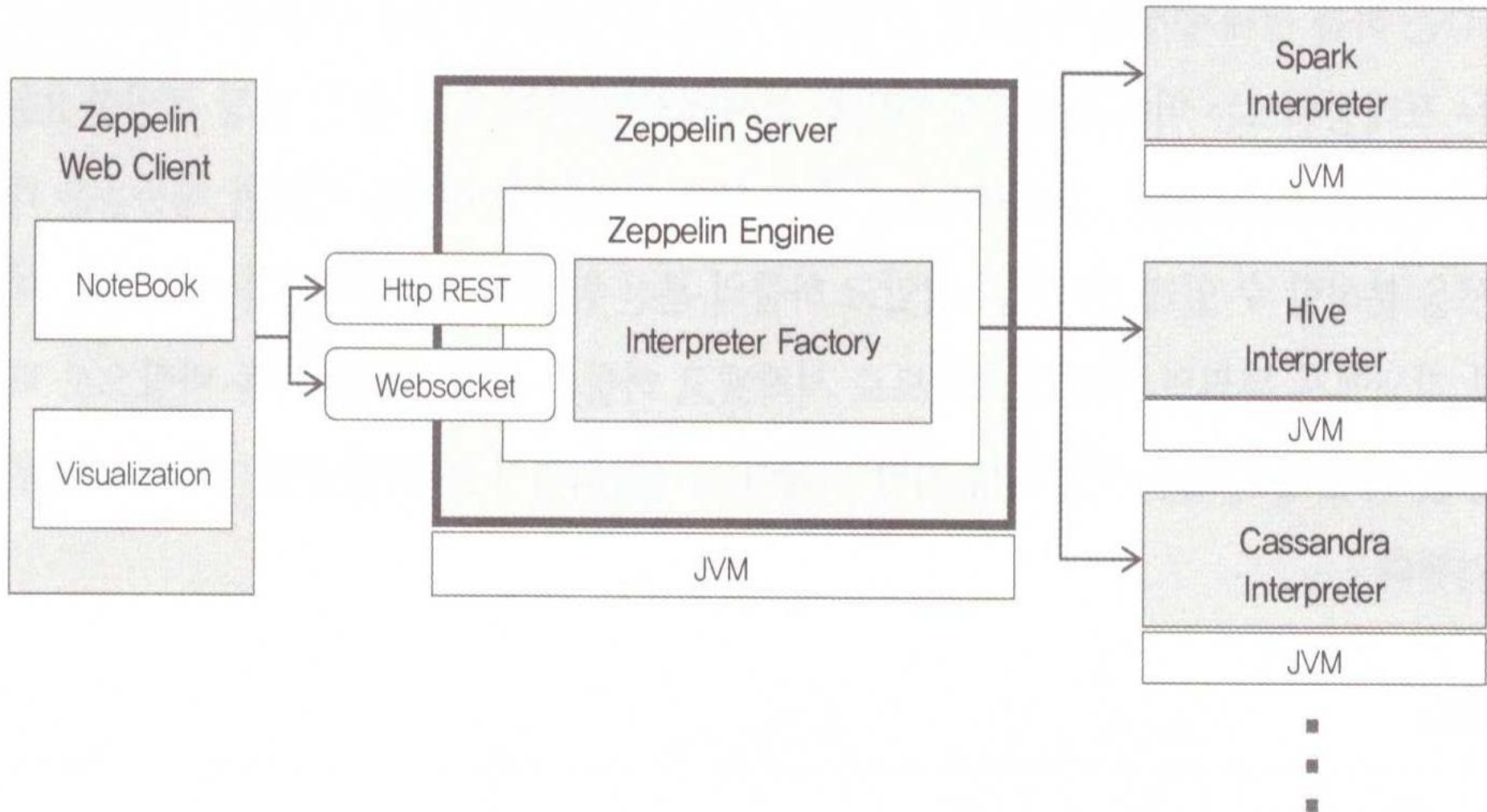
빅데이터 분석 활용 기술 - 제플린(Zeppelin)

➤ 제플린 기본 요소

공식 홈페이지		http://zeppelin.apache.org
주요 구성 요소	NoteBook	웹 상에서 제플린의 인터프리터 언어를 작성하고 명령을 실행 및 관리할 수 있는 UI
	Visualization	인터프리터의 실행 결과를 곧바로 웹 상에서 다양한 시각화 도구로 분석해 볼 수 있는 기능
	Zeppelin Server	NoteBook을 웹으로 제공하기 위한 웹 애플리케이션 서버로서 인터프리터 엔진 및 인터프리터 API 등을 지원
	Zeppelin Interpreter	데이터 분석을 위한 다양한 인터프리터를 제공하며, 스파크, 하이버, JDBC, 셀 등이 있으며 필요 시 인터프리터를 추가 확장
라이선스	Apache	
유사 프로젝트	Jupyter, R	

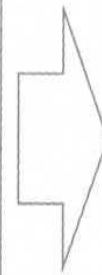
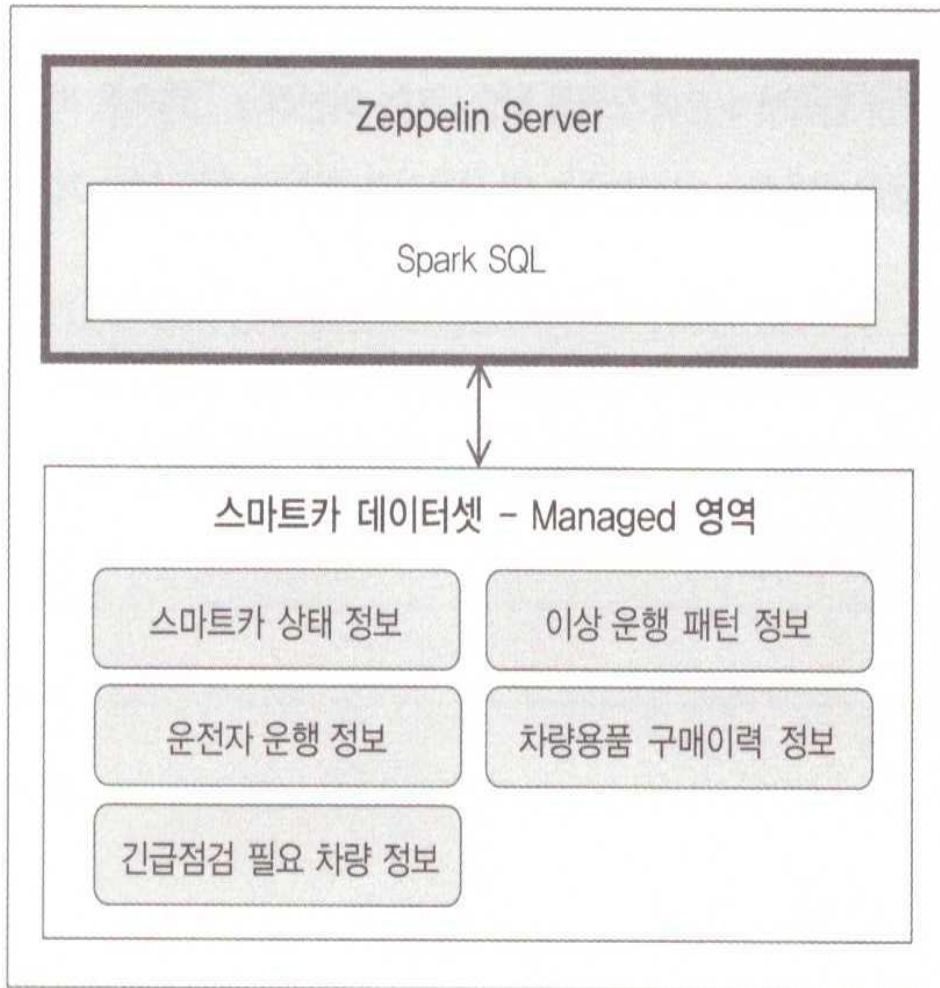
빅데이터 분석 활용 기술 - 제플린(Zeppelin)

➤ 제플린 아키텍처



빅데이터 분석 활용 기술 - 제플린(Zeppelin)

➤ 제플린 활용 방안



- Spark Engine을 이용, 대용량 데이터를 빠르게 Ad-Hoc 분석
- 처리된 결과를 시각화(Table, Line, Pie Chart 등)해서 분석

빅데이터 분석 활용 기술 - 머하웃(Mahout)

➤ 머하웃 소개

- 하둡 생태계에서 머신러닝 기법을 이용해 데이터 마이닝을 수행하는 툴.
- 2008년 검색엔진 루씬의 서브 프로젝트로 시작.
- 하둡의 분산 아키텍처를 바탕으로 텍스트 마이닝, 군집, 분류 등과 같은 머신러닝 기반 기술이 내재화되면서 2010년 4월 아파치 최상위 프로젝트로 승격.

빅데이터 분석 활용 기술 - 머하웃(Mahout)

➤ 머하웃 기본 요소

공식 홈페이지

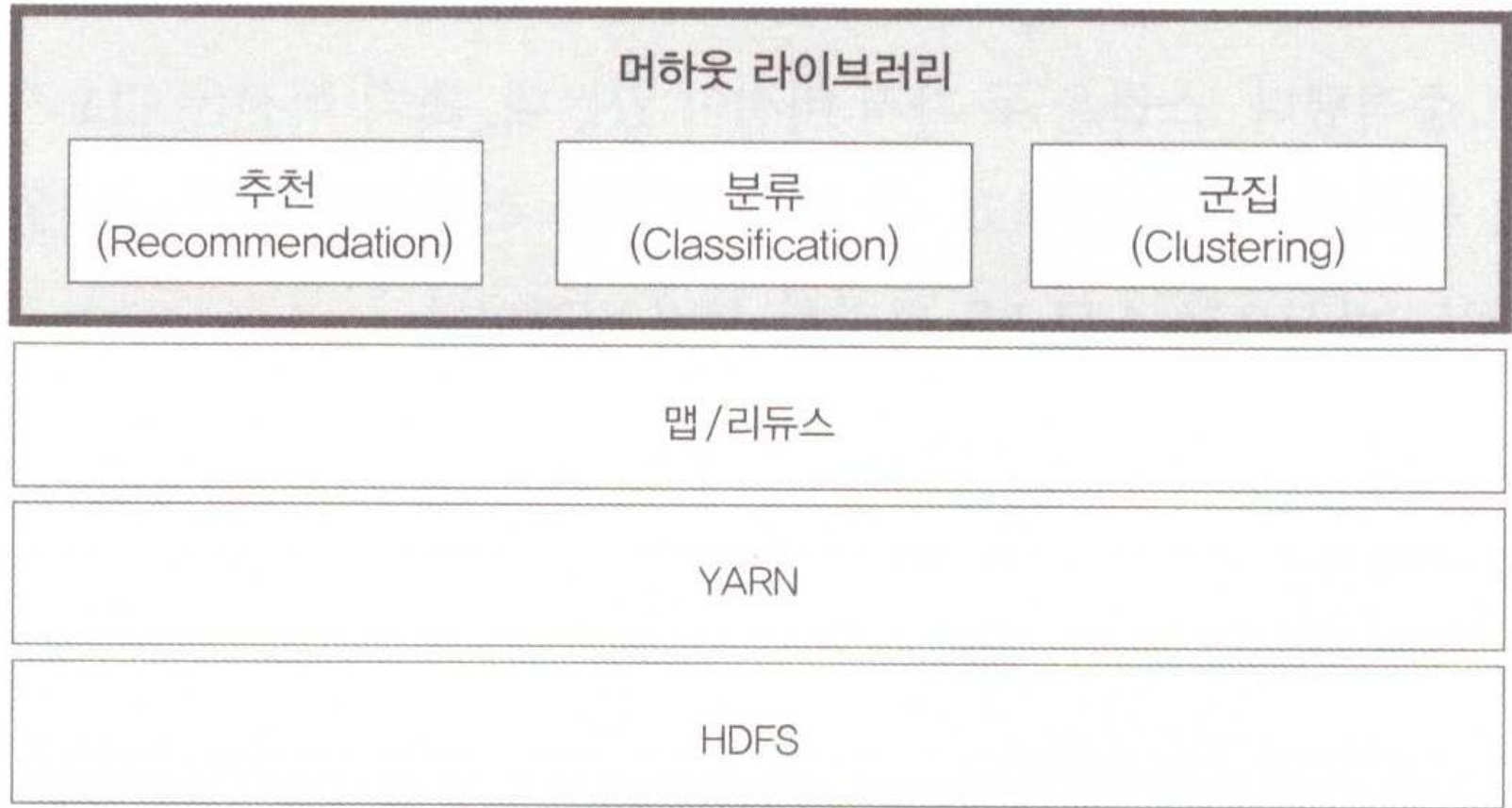


<http://mahout.apache.org>

주요 구성 요소	추천 (Recommendation)	사용자들이 관심을 가졌던 정보나 구매했던 물건의 정보를 분석해서 추천하는 기능으로, 유사한 사용자를 찾아서 추천하는 “사용자 기반 추천”과 항목 간 유사성을 계산해서 추천 항목을 생성하는 “아이템 기반 추천” 등이 존재
	분류 (Classification)	데이터셋의 다양한 패턴과 특징을 발견해 레이블을 지정하고 분류하는 기능으로, 주요 알고리즘으로 나이브 베이지안, 랜덤 포레스트, 로지스틱 회귀 등을 지원
	군집 (Clustering)	대규모 데이터셋에서 새로운 특성으로 데이터의 군집들을 발견하는 기능으로, 주요 알고리즘으로 K-Means, Fuzzy C-Means, Canopy 등을 지원
	감독학습 (Supervised Learning)	학습을 위한 데이터셋을 입력해서 분석 모델을 학습시키는 머신러닝 기법으로, 학습된 분석 모델을 이용해 예측하고 최적화하는 데 사용하고, 분류와 회귀 분석 기법이 이에 해당
	비감독학습 (Unsupervised Learning)	학습 데이터셋을 제공하지 않고 데이터의 특징적인 패턴을 발견하는 머신러닝 기법으로서 사람이 구분 및 그루핑하기 어려운 현상들을 자동으로 그루핑하는 데 사용하며, 군집 기법이 여기에 해당
	라이선스	Apache
유사 프로젝트	R, RapidMiner, Weka	

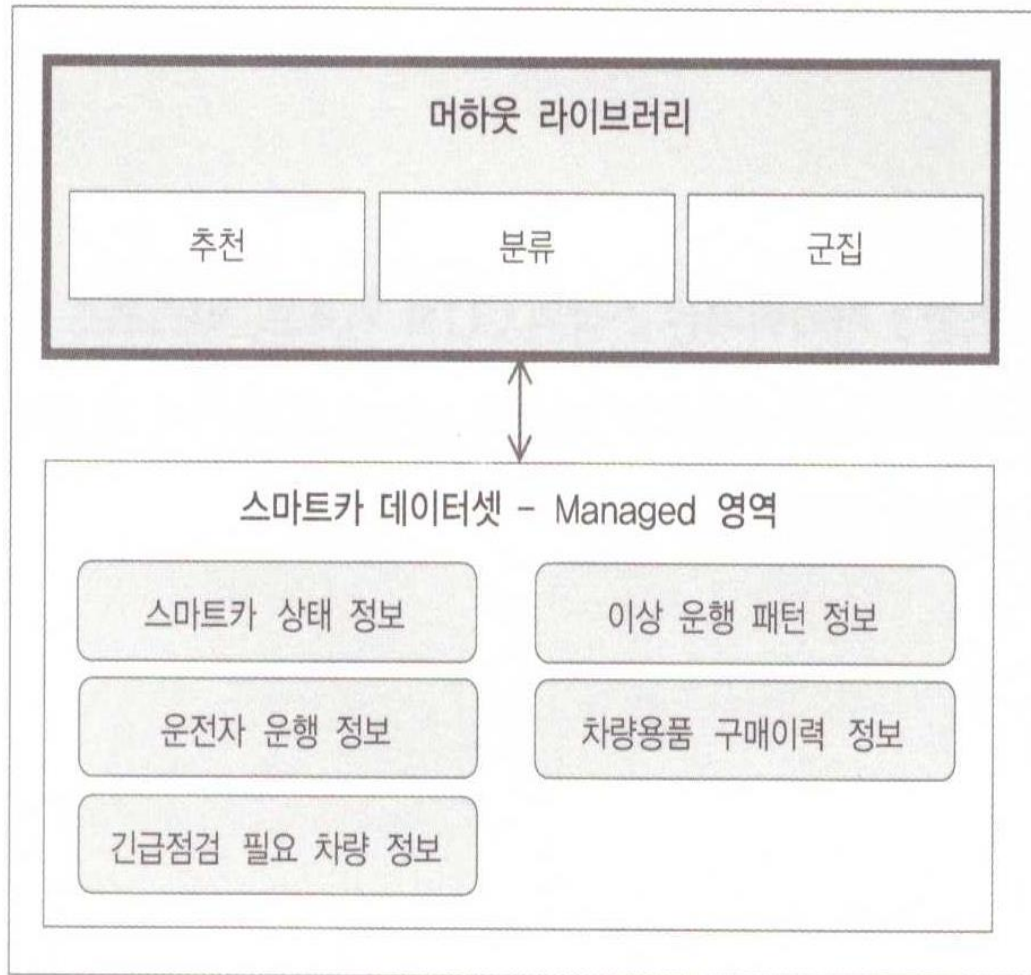
빅데이터 분석 활용 기술 - 머하웃(Mahout)


➤ 머하웃 아키텍처



빅데이터 분석 활용 기술 - 머하웃(Mahout)

➤ 머하웃 활용 방안



- 
- 스마트카 데이터를 활용해 머신러닝의 감독/비감독 학습
 - 차량용품 구매이력 분석으로 스마트카 운전자에게 상품 추천

빅데이터 분석 활용 기술 - 스쿱(Sqoop)

➤ 스쿱 소개

- RDBMS에 있는 데이터를 특별한 전처리 없이 곧바로 HDFS에 적재하거나, 반대로 HDFS에 저장된 데이터를 RDBMS로 제공해야 하는 경우.
- RDBMS와 HDFS 사이에서 데이터를 편리하게 임포트하거나 익스포트 해주는 소프트웨어 기술.
- 2009년에 공개되어 2012년 아파치 최상위 프로젝트로 승격.
- 두 가지 버전 존재
 - 1) 초기 CLI 기반으로 스쿱 명령을 실행하는 스쿱1 클라이언트 버전.
 - 2) 스쿱 서버를 두고 스쿱 클라이언트가 API를 호출하는 방식으로 스쿱1을 확장한 스쿱2 서버 버전.

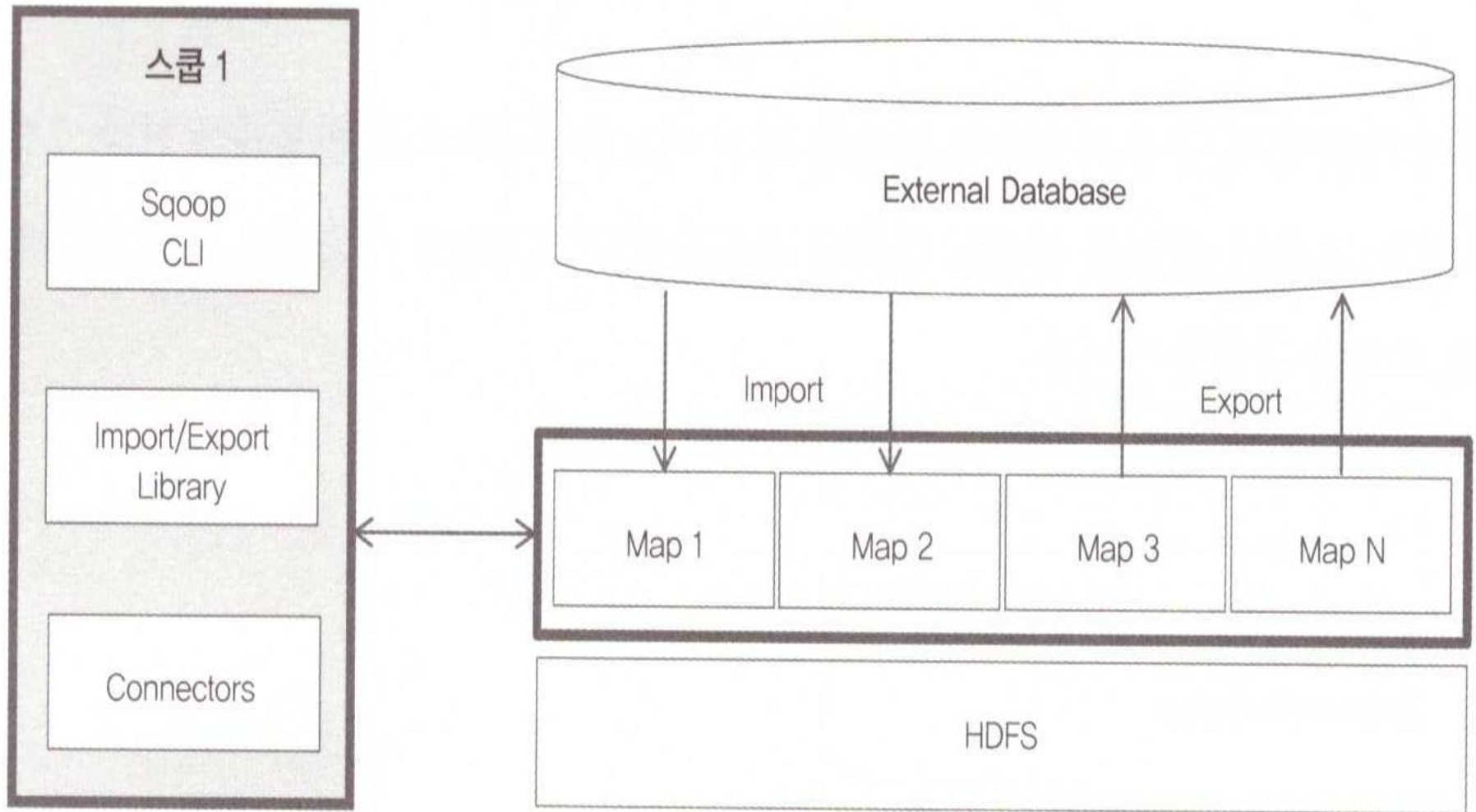
빅데이터 분석 활용 기술 - 스쿱(Sqoop)

➤ 스쿱 기본 요소

공식 홈페이지		http://sqoop.apache.org
주요 구성 요소	Sqoop Client	하둡의 분산 환경에서 HDFS와 RDBMS 간의 데이터 임포트 및 익스포트 기능을 수행하기 위한 라이브러리로 구성
	Sqoop Server	스쿱 2의 아키텍처에서 제공되며, 스쿱 1의 분산된 클라이언트 기능을 통합해 REST API로 제공
	Import/Export	임포트 기능은 RDBMS의 데이터를 HDFS로 가져올때 사용하며, 반대로 익스포트 기능은 HDFS의 데이터를 RDBMS로 내보낼 때 사용
	Connectors	임포트 및 익스포트에서 사용될 다양한 DBMS의 접속 어댑터와 라이브러리를 제공
	Metadata	스쿱 서버를 서비스하는 데 필요한 각종 메타 정보를 저장
라이선스	Apache	
유사 프로젝트	Hiho, Talend, Kettle	

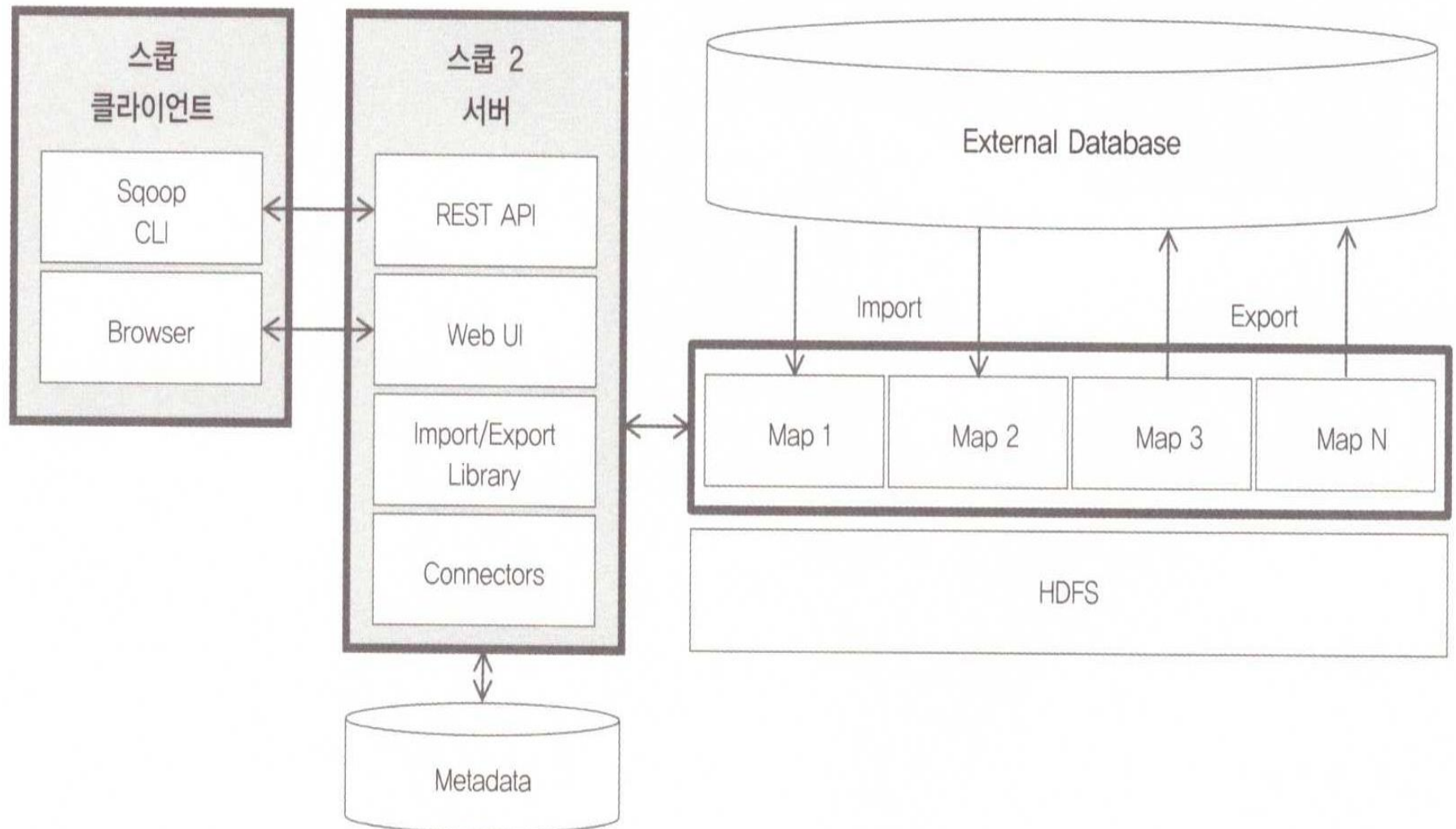
빅데이터 분석 활용 기술 - 스쿱(Sqoop)

➤ 스쿱 아키텍처1



빅데이터 분석 활용 기술 - 스쿱(Sqoop)

➤ 스쿱 아키텍처2



빅데이터 분석 활용 기술 - 스쿱(Sqoop)

➤ 스쿱 활용 방안

