

빅데이터 처리/탐색

# 빅데이터 탐색 소개

## 1. 빅데이터 탐색 개요

빅데이터 탐색에 대한 기본 정의 및 RDBMS 기반의 탐색과 빅데이터 탐색의 차이점을 이해한다.



## 2. 빅데이터 탐색에 활용되는 기술

빅데이터 탐색에 사용할 4가지 기술(하이프, 스파크, 우지, 휴)를 소개하고 각 기술별 주요 기능과 아키텍처, 활용 방안을 알아본다.



## 3. 탐색 파일럿 실행 1단계 - 탐색 아키텍처

스마트카의 빅데이터 탐색과 관련된 요구사항을 구체화하고, 탐색 요구사항을 해결하기 위한 파일럿 아키텍처를 설명한다.



## 4. 탐색 파일럿 실행 2단계 - 탐색 환경 구성

스마트카 탐색 아키텍처를 실제로 설치 및 환경을 구성한다. 하이브, 우지, 휴, 스파크 순으로 설치한다.



## 5. 탐색 파일럿 실행 3단계 - 휴를 이용한 데이터 탐색

휴의 웹 UI를 통해 데이터 탐색 환경을 전반적으로 이해한다. HDFS/HBase/Hive Editor 등을 이용해 하둡에 적재된 스마트카 데이터셋을 탐색한다.



## 6. 탐색 파일럿 실행 4단계 - 데이터 탐색 기능 구현 및 테스트

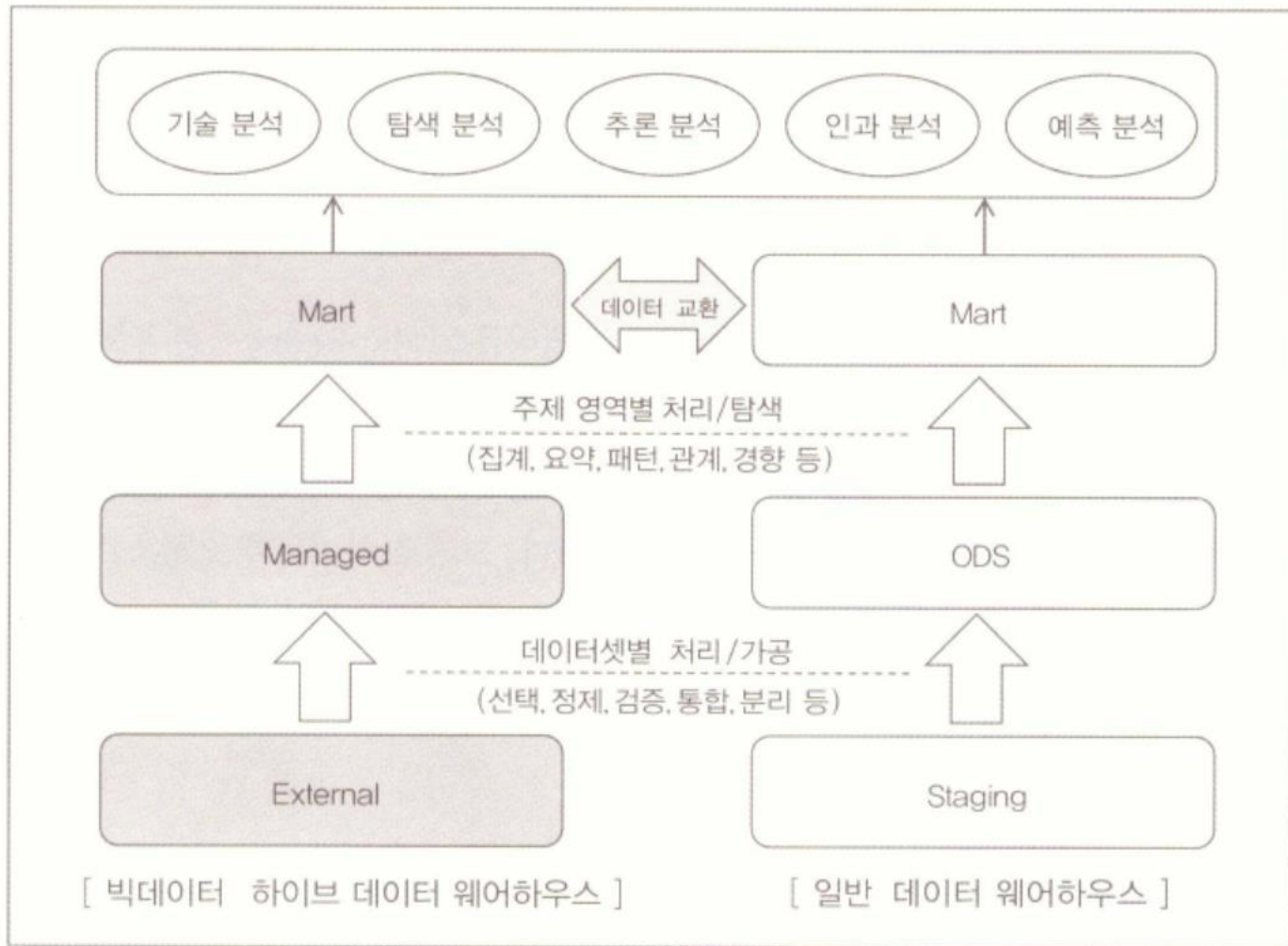
스마트카 데이터셋 탐색 결과를 이용해 5가지 주제 영역에 해당하는 분석 마트를 도출하고, 각 영역별 후처리 워크플로 작업을 한다.

# 빅데이터 처리/탐색 개요

---

- 대용량 저장소에 적재된 데이터를 분석에 활용하기 위해 데이터를 정형화 및 정규화하는 기술.
- 데이터를 통해 가치를 발굴하기 위해서는 데이터를 이해하는 것이 선행되어야 함.
- 적재된 빅데이터를 이해하기 위해 지속적으로 관찰하고 탐색하는 탐색적 분석을 수행.
- 빅데이터 처리 및 탐색 영역은 적재된 데이터를 가공하고 이해하는 단계.
- 탐색적 분석(EDA-Exploratory Data Analysis)
  - 데이터를 이해하는 과정에서의 데이터들의 패턴, 관계, 트렌드 등을 찾는 과정.
  - SQL on Hadoop이 주로 사용.
  - 애드혹(Ad-Hoc) 쿼리로 데이터를 선택, 변환, 통합, 축소 등의 작업을 수행.
- 상당히 많은 시간과 자원이 필요한 단계.
- 덩치 큰 비정형 데이터를 정교한 후처리 작업(필터링, 클리닝, 통합, 분리 등)으로 정형화해서 데이터의 직관성을 확보한 후,
- 업무 도메인에 대한 이해를 바탕으로 충분한 탐색적 분석을 진행했을 때,
- 빅데이터를 통한 미래의 통찰력과 비즈니스 가치의 창출이 가능.
- 탐색 결과는 분석을 위한 기초 데이터로 활용되며, 일련의 처리/탐색, 분석/응용 과정을 거쳐 빅데이터의 DW(Data Warehouse)가 만들어짐.

# 빅데이터 탐색 개요



- 빅데이터 하이브 기반 데이터 웨어하우스 vs. 일반 데이터 웨어하우스

# 빅데이터 Data Warehouse

---

## ➤ External 영역

- 전처리(수집/적재)와 후처리(탐색/분석)가 만나서 데이터를 서로 공유하는 영역.
- 원천 데이터의 형식을 최대한 유지.

## ➤ Managed 영역

- 처리/가공 단계를 거친 External의 데이터셋이 전달되어 옴.
- 데이터의 주제 영역별 처리/탐색 과정.

## ➤ Mart 영역

- 현황 분석 모형 : 빅데이터 마트 모델을 통합, 요약, 집계 등을 리포팅.
- 고급 분석 모형 : 데이터의 패턴과 트렌드를 분석해 미래를 예측.

## ➤ 하이브리드 DW

- 빅데이터 마트와 일반 마트가 데이터를 서로 교환 가능.
- 대규모 거래 및 실시간성 데이터의 처리는 빅데이터 DW가, 온라인성 업무 및 마스터 데이터는 기존 DW가 처리하면서 서로의 단점을 보완.

# 빅데이터 탐색에 활용하는 기술 - Hive

---

## ➤ Hive 소개

- 하둡 초창기에 적재된 데이터를 탐색/분석하기 위한 도구로 MapReduce를 주로 이용.
- MapReduce는 복잡도가 높은 프로그래밍 기법이 필요했고, 이는 업무 분석가 및 관리자들에게 빅데이터에 접근하는 것을 어렵게 만들.
- 이를 해결하기 위해 페이스북에서 SQL과 매우 유사한 방식으로 하둡 데이터에 접근성을 높인 Hive 개발.
- 오픈 소스로 공개되면서 2016년 2월 하이브 2.0이 릴리스.
- 빅데이터의 가장 대표적인 SQL on Hadoop 제품으로 자리 잡음.
- 테이블 형태의 논리적인 뷰도 제공하며, DW(data warehouse)를 구축하기 위한 용도로도 활용.

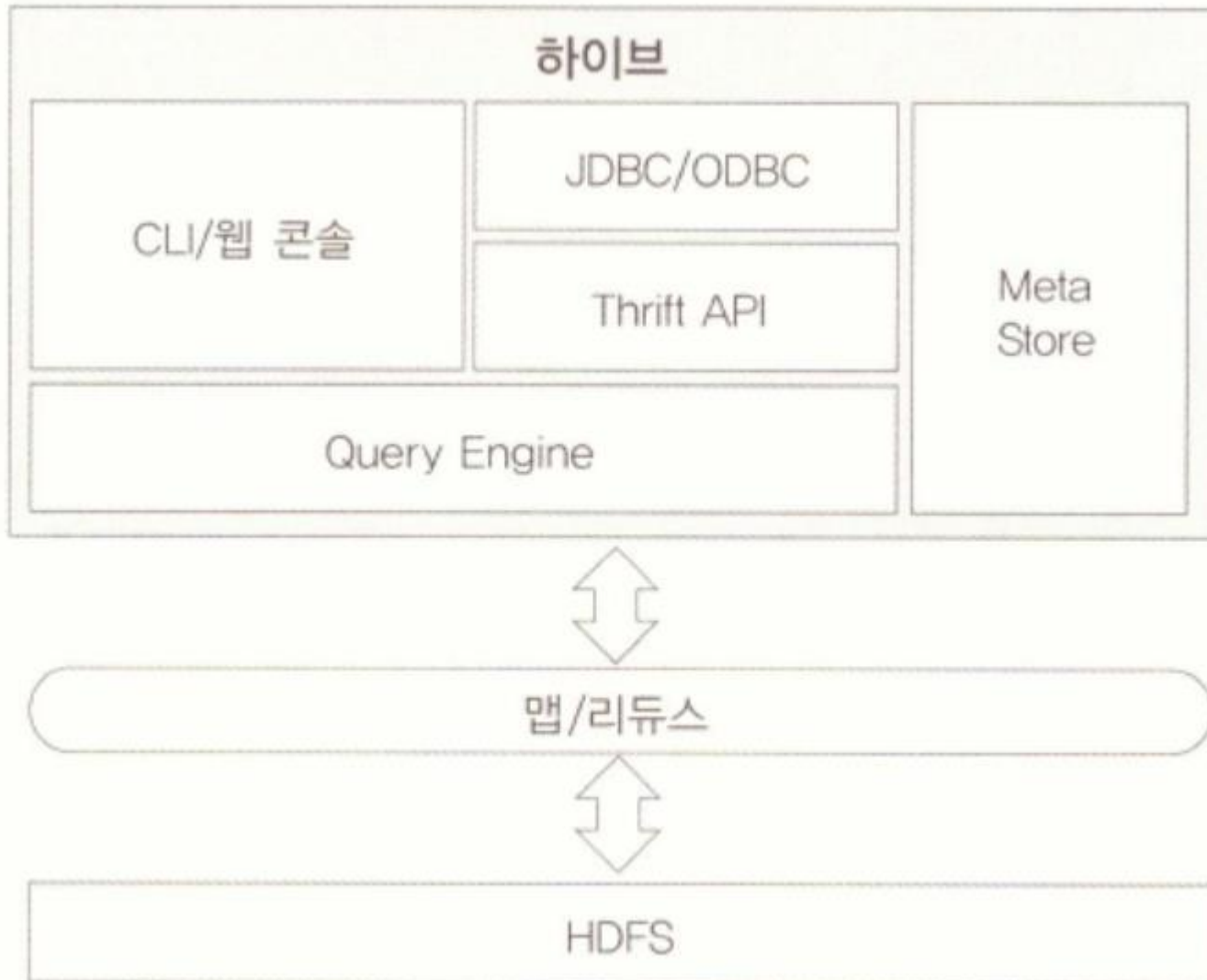
# 빅데이터 탐색에 활용하는 기술 - Hive

## ➤ Hive 기본 요소

공식 홈페이지		<a href="http://hive.apache.org">http://hive.apache.org</a>
주요 구성 요소	CLI	사용자가 하이브 쿼리를 입력하고 실행할 수 있는 인터페이스(Hive Server1 기반의 CLI와 Hive Server2 기반의 Beeline이 있음)
	JDBC/ODBC Driver	하이브의 쿼리를 다양한 데이터베이스와 연결하기 위한 드라이버를 제공
	Query Engine	사용자가 입력한 하이브 쿼리를 분석해 실행 계획을 수립하고 하이브 QL을 맵리듀스 코드로 변환 및 실행
주요 구성 요소	MetaStore	하이브에서 사용하는 테이블의 스키마 정보를 저장 및 관리하며, 기본적으로 더비 DB(Derby DB)가 사용되나 다른 DBMS(MySQL, PostgreSQL 등)로 변경 가능
라이선스	Apache	
유사 프로젝트	Impala, Tajo, Spark-SQL, Presto	

# 빅데이터 탐색에 활용하는 기술 - Hive

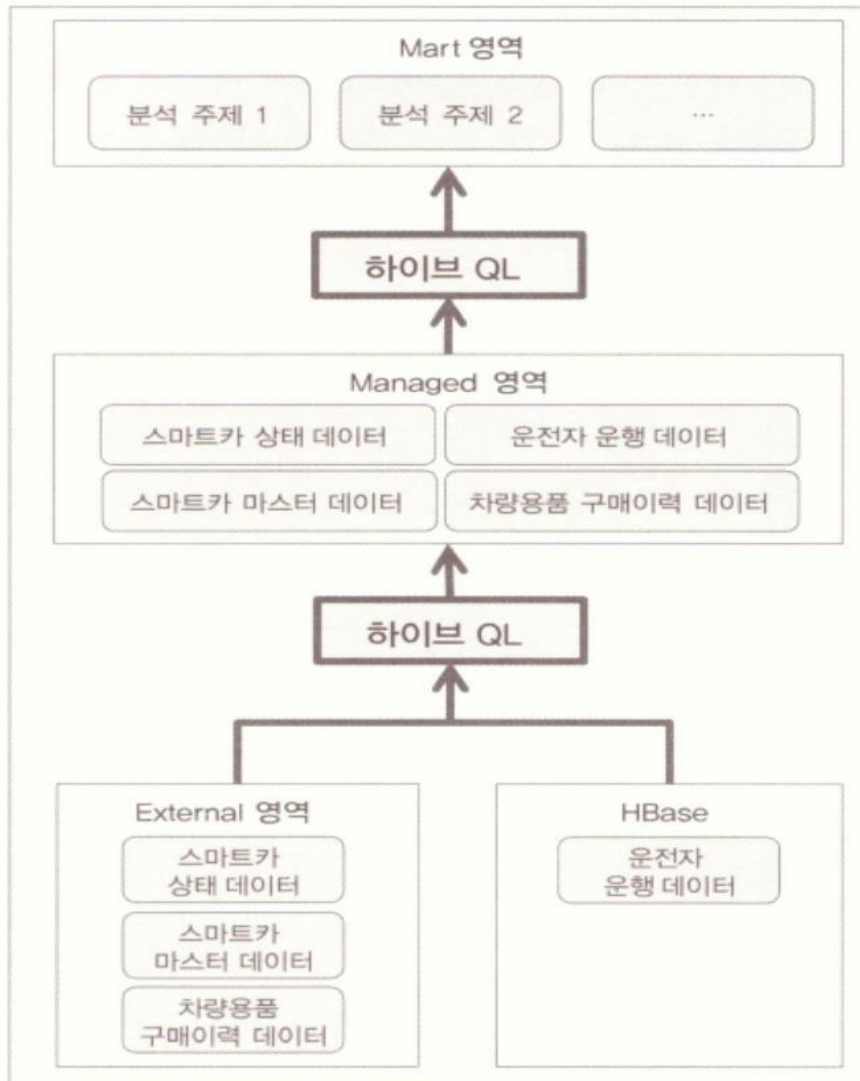
## ➤ Hive 아키텍처





# 빅데이터 탐색에 활용하는 기술 - Hive

## ➤ Hive 활용 방안



- 빅데이터 기반 스마트카 데이터 웨어하우스 구축
- 스마트카 및 운행 데이터에 대한 Cleansing, Filtering, Transformation 작업

# 빅데이터 탐색에 활용하는 기술 - Spark

---

## ➤ Spark 소개

- 맵리듀스 코어를 그대로 사용하는 하이브는 성능면에서 만족스럽지 못함.
- 그로 인해 반복적인 대화형 연산 작업에서는 하이브가 적합하지 못함.
- 이 단점을 극복한 고성능 인메모리 분석.
- 하둡과 유사한 클러스터 기반의 분산 처리 기능을 제공하는 오픈소스 프레임워크.
- UC 버클리의 AMPLab에서 2009년 개발, 2010년 오픈 소스로 공개.
- 2013년 6월 아파치 재단으로 이관되어 최상위 프로젝트.
- 최근 빅데이터 분야에서 가장 핫한 기술 중 하나.
- 데이터 가공 처리를 인메모리에서 수행함으로써 대용량 데이터 작업에도 빠른 성능을 보장.
- 하둡과 하이브를 비롯한 기존의 여러 솔루션과의 연동을 지원하고 마이크로배치 방식의 실시간 처리 및 머신러닝 라이브러리를 비롯해 빅데이터 처리와 관련된 다양한 라이브러리를 지원.

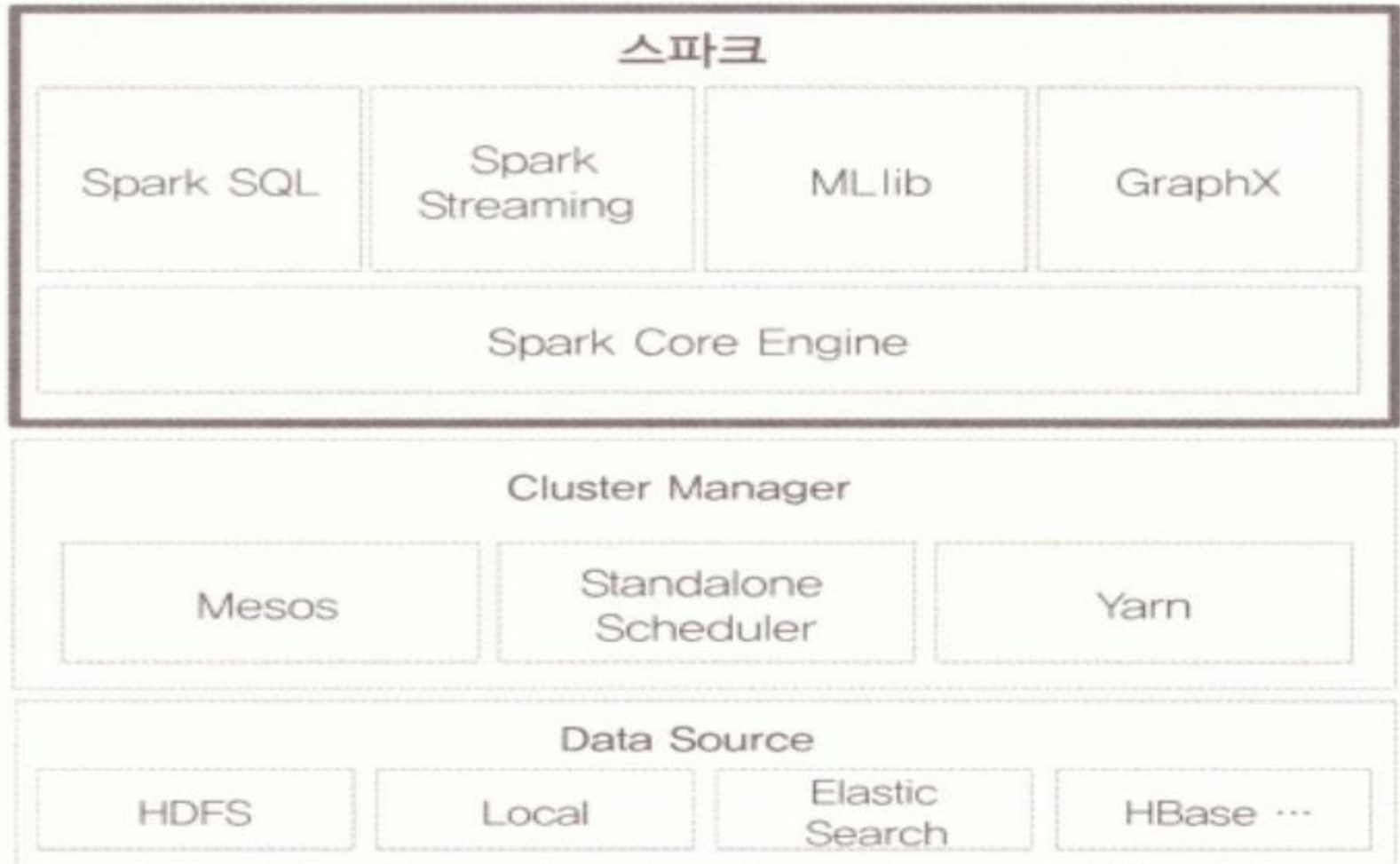
# 빅데이터 탐색에 활용하는 기술 - Spark

## ➤ Spark 기본 요소

공식 홈페이지		<a href="http://spark.apache.org">http://spark.apache.org</a>
주요 구성 요소	Spark RDD	스파크 프로그래밍의 기초 데이터셋 모델
	Spark Driver / Executors	Driver는 RDD 프로그램을 분산 노드에서 실행하기 위한 Task의 구성, 할당, 계획 등을 수립하고, Executor는 Task를 실행 관리하며, 분산 노드의 스토리지 및 메모리를 참조
	Spark Cluster Manager	스파크 실행 환경을 구성하는 클러스터 관리자로 Mesos, YARN, Spark Standalone이 있음
	Spark SQL	SQL 방식으로 스파크 RDD 프로그래밍을 지원
	Spark Streaming	스트리밍 데이터를 마이크로타임의 배치로 나누어 실시간 처리
	Spark MLlib	스파크에서 머신러닝 프로그래밍(군집, 분류, 추천 등)을 지원
	Spark GraphX	다양한 유형의 네트워크(SNS, 하이퍼링크 등) 구조 분석을 지원
라이선스	Apache	
유사 프로젝트	Impala, Tajo, Tez	

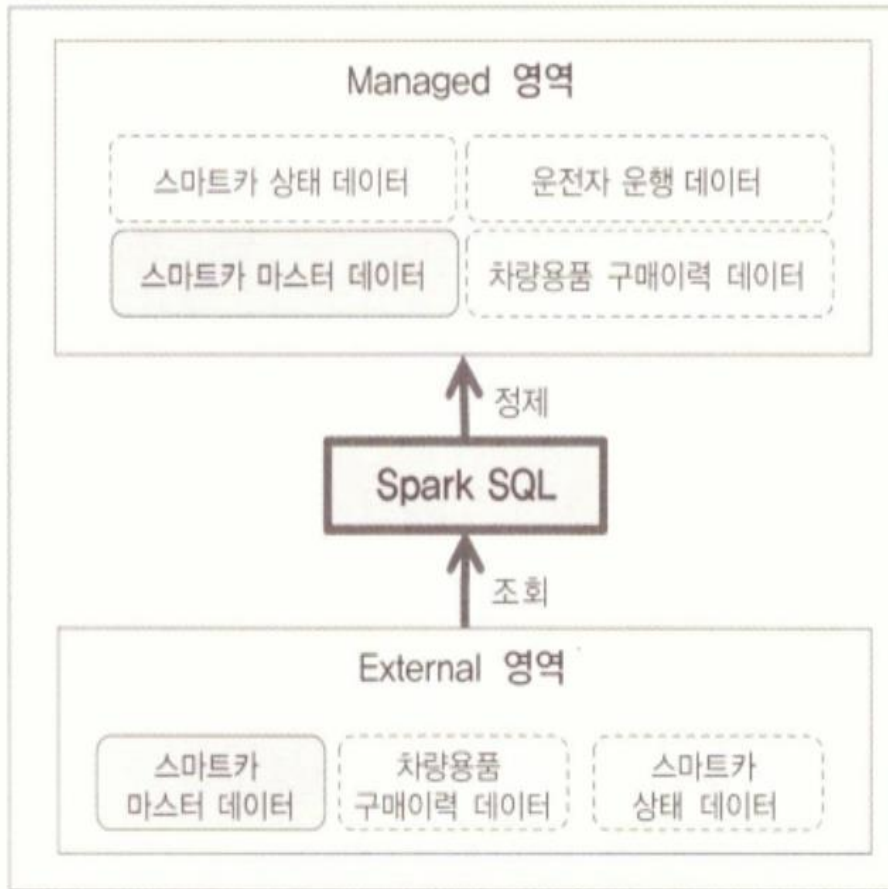
# 빅데이터 탐색에 활용하는 기술 - Spark

## ➤ Spark 아키텍처



# 빅데이터 탐색에 활용하는 기술 - Spark

## ➤ Spark 활용 방안



- External에 적재된 스마트카 마스터 데이터 조회
- 마스터 데이터 정제 처리 후, Managed에 적재

# 빅데이터 탐색에 활용하는 기술 - Oozie


---

## ➤ Oozie 소개

- 수집 및 적재된 수백 개 이상의 데이터셋을 대상으로 다양한 후처리 job이 데이터 간의 의존성과 무결성을 유지하며 복잡하게 실행됨.
- 반복적이면서 복잡한 후처리 job을 처리하기 위해 방향성 있는 비순환 그래프(DAG:Direct Acyclic Graph)로 정의해서 job에 시작, 처리, 분기, 종료점 등의 액션(Action)으로 구성하는 워크플로(workflow)가 필요.
- 위의 필요성에 의해 만들어진 것이 아파치 우지.
- 2008년 야후에서 개발, 2010년 오픈 소스로 공개.
- 2012년 아파치 최상위 프로젝트로 승격되어, 2015년 버전 4.2까지 릴리즈.

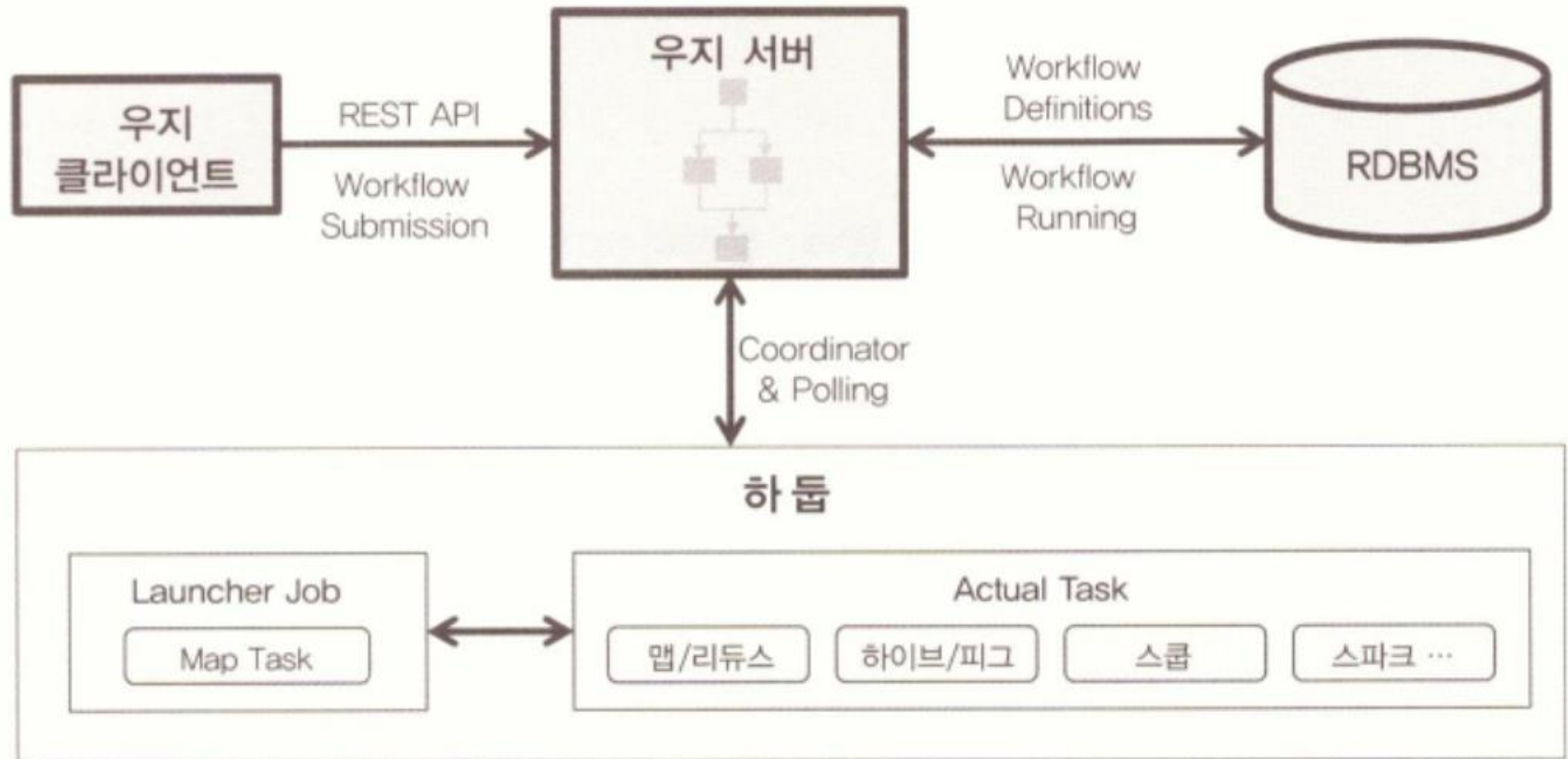
# 빅데이터 탐색에 활용하는 기술 - Oozie

## ➤ Oozie 기본 요소

공식 홈페이지		<a href="http://oozie.apache.org">http://oozie.apache.org</a>
주요 구성 요소	Oozie Workflow	주요 액션에 대한 작업 규칙과 플로우를 정의
	Oozie Client	워크플로를 Server에 전송하고 관리하기 위한 환경
주요 구성 요소	Oozie Server	워크플로 정보가 잡으로 등록되어 잡의 실행, 중지, 모니터링 등을 관리
	Control 노드	워크플로의 흐름을 제어하기 위한 Start, End, Decision 노드 등의 기능을 제공
	Action 노드	잡의 실제 수행 태스크를 정의하는 노드로서 하이브, 피그, 맵리듀스 등의 액션으로 구성
	Coordinator	워크플로 잡을 실행하기 위한 스케줄 정책을 관리
라이선스	Apache	
유사 프로젝트	Azkaban, Cascading, Hamake, Airflow	

# 빅데이터 탐색에 활용하는 기술 - Oozie

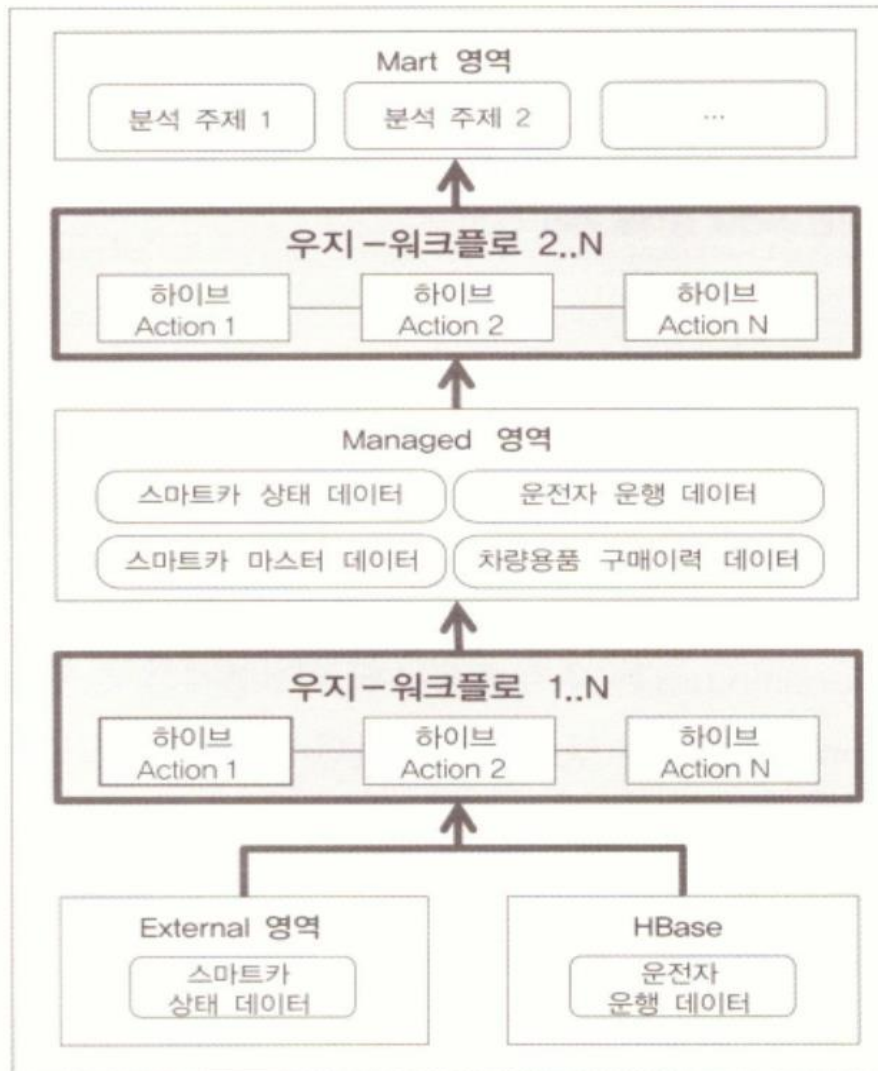
## ➤ Oozie 아키텍처





# 빅데이터 탐색에 활용하는 기술 - Oozie

## ➤ Oozie 활용 방안



- 스마트카의 빅데이터 웨어하우스 구축을 위한 단계별 워크플로 기능 제공
- 워크플로를 주기적으로 실행 및 관리하기 위한 Coordinator 기능 제공

# 빅데이터 탐색에 활용하는 기술 - Hue

---

## ➤ Hue 소개

- 빅데이터 탐색/분석은 반복적인 작업이면서 그 과정에서 많은 도구들이 활용.
- 하둡 기반의 하이브, 피그, 우지, 스쿱 등 알아야 할 기술 요소가 지나치게 많아 업무 담당자 또는 데이터 분석가들이 직접 사용하기에는 많은 어려움.
- 빅데이터 기술이 성숙해지면서 이러한 복잡도를 숨기고 접근성을 높인 소프트웨어들이 만들어짐.
- 그중 하나가 클라우데라에서 만든 것이 Hue.
- Hue는 다양한 하둡의 에코시스템의 기능들을 웹 UI로 통합 제공.
- 오픈 소스로 깃허브에 공개, 2016년 공식 사이트에서 3.90 버전까지 릴리즈.

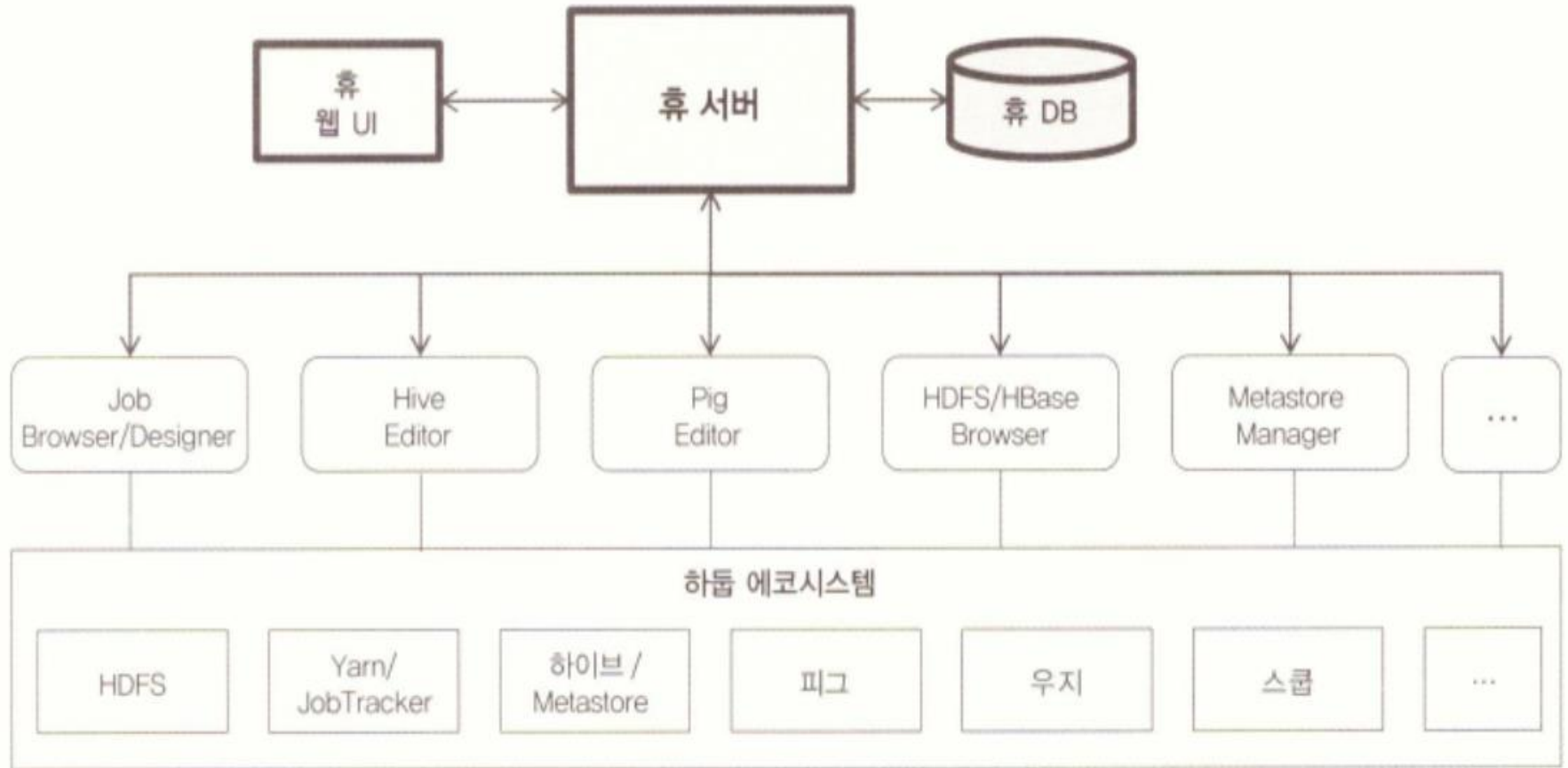
# 빅데이터 탐색에 활용하는 기술 - Hue

## ➤ Hue 기본 요소

공식 홈페이지		<a href="http://gethue.com">http://gethue.com</a>
주요 구성 요소	Job Designer	우지의 워크플로 및 Coordinator를 웹 UI에서 디자인
	Job Browser	등록한 잡의 리스트 및 진행 상황과 결과 등을 조회
	Hive Editor	하이프 QL을 웹 UI에서 작성, 실행, 관리
	Pig Editor	피그 스크립트를 웹 UI에서 작성, 실행, 관리
	HDFS Browser	하둡의 파일시스템을 웹 UI에서 탐색 및 관리
	HBase Browser	HBase의 HTable을 웹 UI에서 탐색 및 관리
라이선스	Apache	
유사 프로젝트	NDAP, Flamingo, Ambari	

# 빅데이터 탐색에 활용하는 기술 - Hue

## ➤ Hue 아키텍처



# 빅데이터 탐색에 활용하는 기술 - Hue

## ➤ Hue 활용 방안

