6.2 빅데이터 탐색에 활용되는 기술

하이브

공식 홈페이지		http://hive.apache.org
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 QL(Query Language)을 맵리듀스 코드로 변환 및 실행
	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	

6.2 빅데이터 탐색에 활용되는 기술

하이브 아키텍처

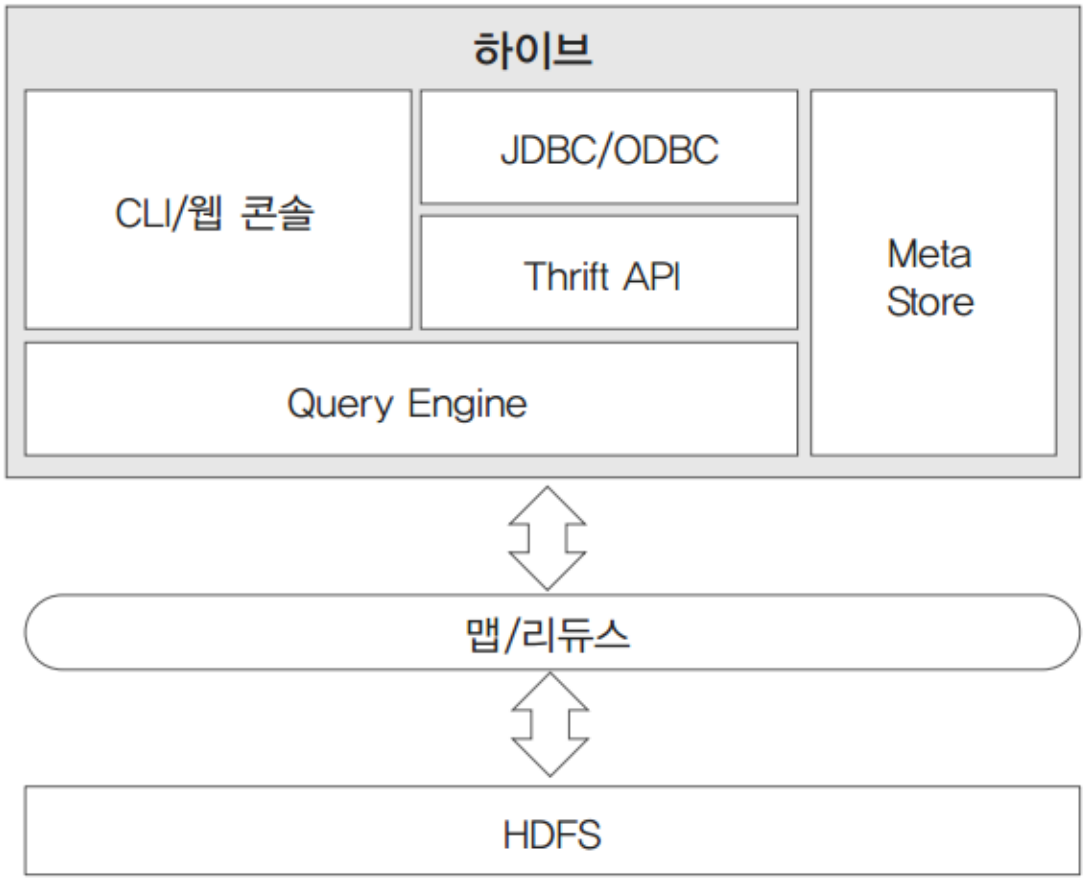
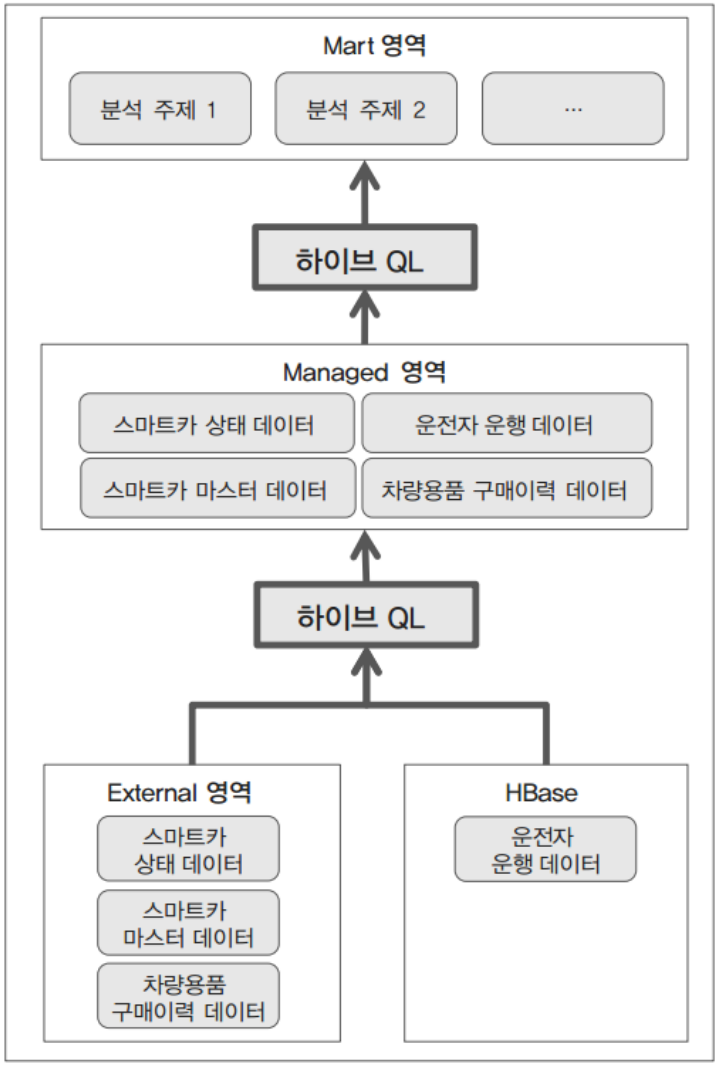


그림 6.4 하이브 아키텍처

6.2 빅데이터 탐색에 활용되는 기술

하이브 활용 방안



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

Tip _ 하이브 External? or Managed?

하이브의 데이터 아키텍처는 External 영역과 Managed 영역으로 구분된다. 데이터 제어와 관리에서 두 영역에 차이가 있는데, 스키마에 디펜던시가 작은 영역이 External 영역이고 그 반대가 Managed 영역이다. 비정형의 3V 데이터가 쌓이는 빅데이터 레이크 영역을 External에, 여기서 정형화한 빅데이터 웨어하우스나 마트 영역은 Managed 영역에 만든다(참고로 마트 영역은 성능과 연동 등을 고려해 하이브가 아닌 별도의 RDBMS에 구성하기도 한다).

Tip _ 피그란?

하둡 에코시스템 가운데 하이브와 유사한 목적으로 맵리듀스의 복잡성을 해결하기 위한 피그(Pig)라는 프로젝트가 있다. 피그는 SQL 대신 피그 라틴(Pig Latin)이라는 언어를 제공해서 하이브보다는 절차적인 요소가 많이 사용되는 특징이 있다. 하이브가 테이블을 기반으로 한 데이터 가공, 적재, 탐색 등에 최적화됐다면 피그는 HDFS 파일에 직접 접근해 다양한 데이터 처리 함수와 제어문으로 복잡한 데이터 파이프라인을 처리하는 데 적합하다.

피그 설치 및 활용법을 유튜브에 올려 놓았으니 참고하기 바란다.

- 실무로 배우는 빅데이터 기술(확장편): https://bit.ly/bigdata2nd