Introduction to Hive

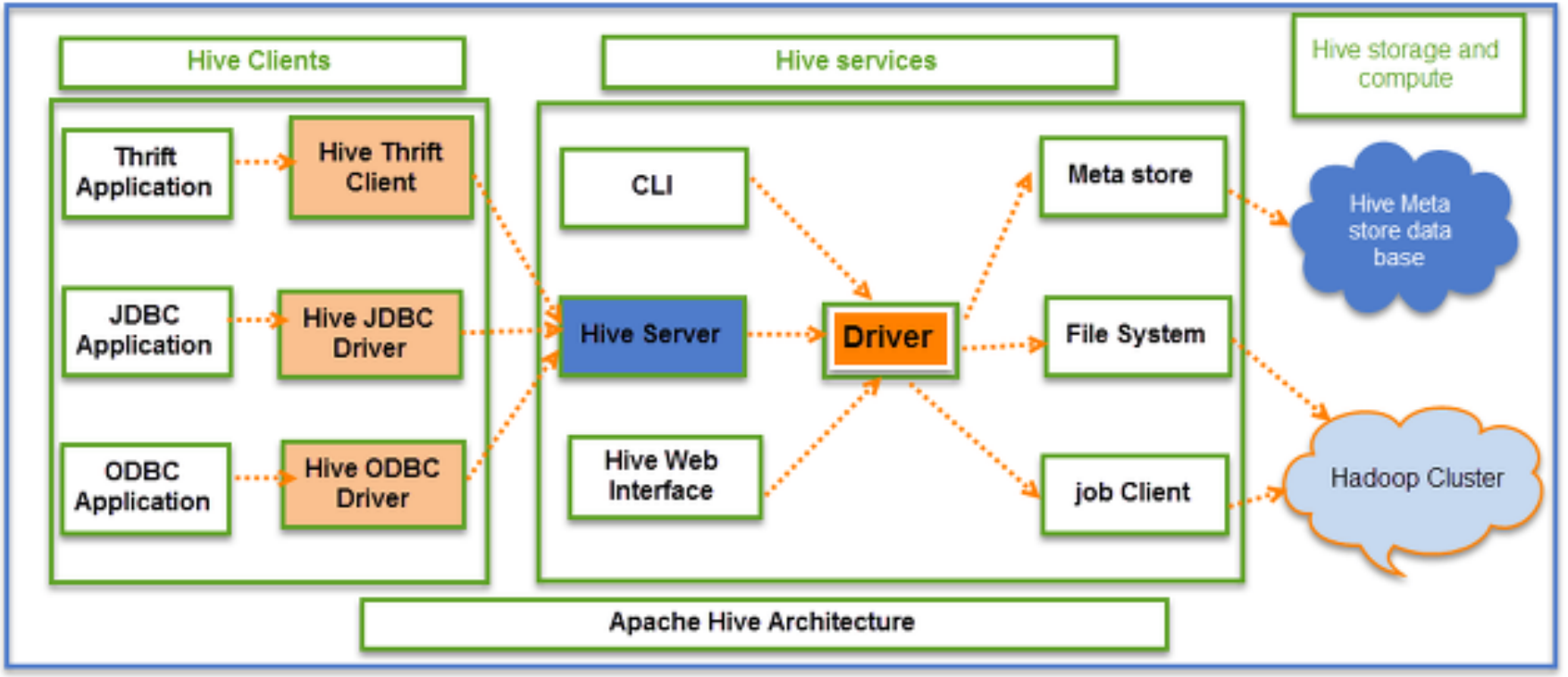
Hive?

- 하둡에 저장된 데이터를 쉽게 처리할 수 있도록 개발된 데이터웨어하우스 패키지

- Facebook에서 개발 → Apache Project 등록

- SQL과 OLAP 솔루션에 익숙한 데이터 분석 업무 전문가들이 Java와 같은 프로그래밍 언어를 사용하지 않고도 Hadoop 시스템에서 쉽게 작업할 수 있도록 지원

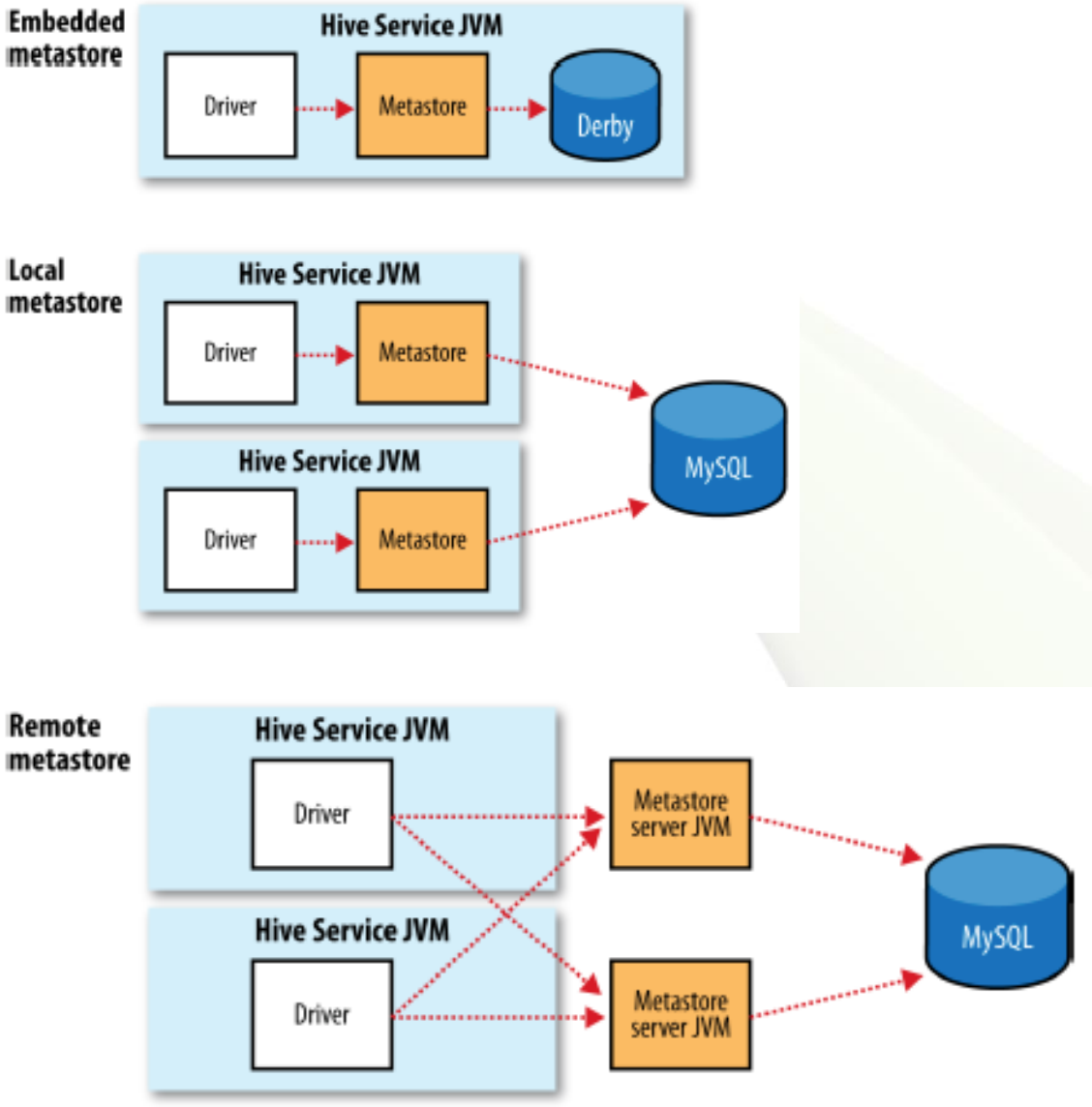
Hive Architecture



Hive 구성 요소

구성 요소	설명
Hive 서비스	하이브 쉘에 대한 명령행 인터페이스 (기본 서비스) 하이브 서버 : 다른 언어로 개발된 클라이언트와 연동하는 서비스
클라이언트	하이브 시스템을 통해 하둡 기반 데이터를 처리하는 애플리케이션 쓰리프트 서비스 제공 → 쓰리프트 지원 언어는 하이브 연결 가능 JDBC 드라이버 및 ODBC 드라이버 지원
메타스토어	데이터 구조를 저장하는 저장소 • 단일 사용자인 경우 apache derby 사용 (default, 내장형) • 다중 사용자인 경우 RDBMS를 사용해서 메타스토어 구성
드라이버	클라이언트가 요청한 HIVE-QL 구문 해석 실행 계획 작성 및 최적화 수행 맵리듀스 잡 실행

Hive Metastore



Hive QL

- Hive에서 사용하는 명령어 구문으로 SQL과 유사한 형식
  - SQL-92, MySQL, Oracle 등의 구문 혼합

- 대부분 SQL과 유사하지만 Hadoop 시스템의 특성으로 인해 제약 사항 있음
  - HDFS는 수정에 제약이 있기 때문에 INSERT, UPDATE, DELETE 등의 구문 사용 제한
  - 제한된 트랜잭션 지원
  - 서브쿼리는 FROM, WHERE, HAVING 절에서 지원
  - 뷰는 읽기 전용으로 지원

구문 종류	지원 구문
DDL	CREATE DATABASE CREATE TABLE ALTER TABLE SHOW TABLE DESCRIBE …
QUERY	SELECT GROUP BY JOIN ORDER BY …
DML	LOAD TABLE INSERT …

Hive QL 자료형

- Numeric Types
  - TINYINT (1-byte 부호 있는 정수)
  - SMALLINT (2-byte 부호 있는 정수)
  - INT/INTEGER (4-byte 부호 있는 정수)
  - BIGINT (8-byte 부호 있는 정수)
  - FLOAT (4-byte 단정밀도 부동소수점)
  - DOUBLE (8-byte 배정밀도 부동소수점)
  - DOUBLE PRECISION (alias for DOUBLE)
  - DECIMAL (임의 정밀도 부호 있는 숫자)
  - NUMERIC (same as DECIMAL)

- Date/Time Types
  - TIMESTAMP (10억분의 1초 정밀도)
  - DATE (날짜)
  - INTERVAL (시간 간격)

- String Types
  - STRING (길이 제한 없음)
  - VARCHAR (가변 길이)
  - CHAR (고정 길이)

- Misc Types
  - BOOLEAN
  - BINARY (바이트 배열)

테이블

- 관리 테이블
  - 하이브가 데이터를 직접 관리
  - 데이터를 하이브가 관리하는 Data Warehouse 디렉터리로 복사
  - CREATE TABLE …
- 외부 데이블
  - 하이브가 관리하는 Data Warehouse 외부에 데이터 저장
  - CREATE EXTERNAL TABLE …

파티션과 버킷

- 파티션
  - 테이블의 데이터를 날짜와 같은 파티션 컬럼을 기반으로 분할해서 관리
  - 데이터 조회 성능 향상에 기여
  - CREATE TABLE … PARTITION BY 구문 사용

- 버킷
  - 효율적인 조회 가능
  - 버킷 컬럼 해시를 기준으로 지정된 개수의 파일로 분리해서 데이터 관리
    - 파티션 내부의 분리된 파일
  - CREATE TABLE … CLUSTERED BY 구문 사용

데이터 저장 포맷

- 파일 포맷
  - 파일에 레코드를 저장할 때의 인코딩 방식
  - CREATE TABLE … STORED AS 구문 사용

- 종류
  - 텍스트 파일
  - 시퀀스 파일
  - RC 파일
  - ORC 파일
  - 파케이
  - …

예제 데이터 구조

1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number

11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes

21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes