

## 자바쌤~!

<https://cafe.naver.com/javassem>

검색

63기-자바 &amp; 빅데이터 &gt;

## [ 하이브 ] 개념



자바쌤 1:1 채팅

2020.06.23. 16:17 조회 30

댓글 0 URL 복사



## 1. HIVE 개요

자바 기반의 매퍼/리듀스 프로그래밍이 너무 어려워서 나온 것들이 Pig이다.

이 Pig도 Pig Latin 언어를 이용한 스크립트 언어이다.

그래서 기존 DB분야 사람들에게는 익숙한 SQL을 사용하도록 만들어진 것이 HIVE이다.

( Hive QL - query language )



## 1. HIVE 개요

SQL과 유사한 HiveQL 사용

매퍼/리듀스 프로그램 대신 쿼리 인터페이스 서비스 제공

## 카페정보 나의활동



매니저 자바쌤  
since 2013.03.15.  
카페소개

새싹2단계

437

초대하기

즐거찾는 멤버

64명

게시판 구독수

4회

우리카페업 수

0회

카페 글쓰기

카페 채팅

## 즐거찾는 게시판

전체글보기

4,052

## KOSMO

63기-자바 &amp; 빅데이터

## 팁

- 심심문제
- 기타등등
- JAVA
- Oracle/Mysql
- JSP
- WebUI
- 파이썬
- Spring
- 리눅스
- 하둡+eco
- 프로젝트

## 공유해요

- 알고리즘도전
- 기술면접 준비하기
- 실존 면접
- 기회는 준비된자에게
- 프로젝트 팁

## 쉬어가기

## 청년아카데미

## 전문가과정

공지사항  
카페회칙

카페탈퇴하기

쿼리 실행시 맵리듀스로 진화되어 결과를 생성

RDBMS 개념처럼 쿼리를 수행하기에 비정형 데이터는 적합하지 않음

## 2. Hive Architecture

### [ Hive CLI ]

cli : command line interface

#### 1. Hive

· 하이브 메타 스토어 및 드라이버에 직접 액세스

\$ hive

```
hive> exit;
```

#### 2. Beeline

· hiveserver2 API를 사용하여 더욱 안전함.

· 보다 나은 인증 및 권한부여 제공

\$ beeline

```
beeline> !connect jdbc:hive2://
Enter username for jdbc:hive2://: hive
Enter password for jdbc:hive2://: hive
0: jdbc:hive2://> !quit
```

## 2. Hive Architecture

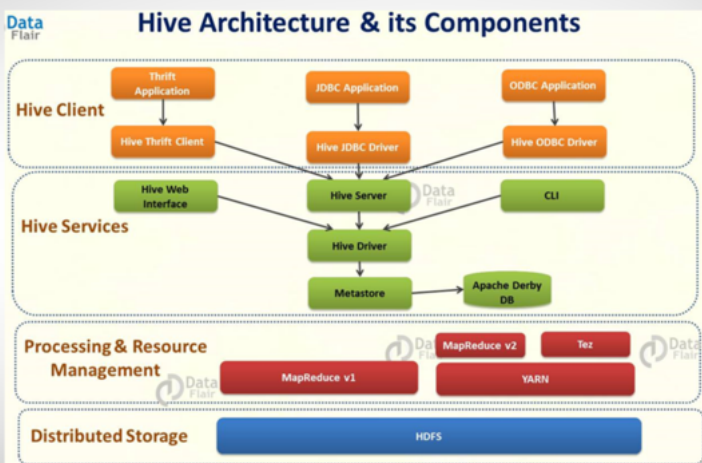
### 1. 하이브 클라이언트

- JDBC 기반 응용 프로그램 지원
- Thrift 기반 응용 프로그램 지원  
(Thrift도 통신 프로토콜의 일종으로 데이터를 주고 받을 수 있다)
- ODBC 기반 응용 프로그램 지원

### 2. 하이브 서비스

- Hive Server
- CLI / Hive Web Interface
- Hive Driver
- Metastore / Apache Derby DB

## 2. Hive Architecture



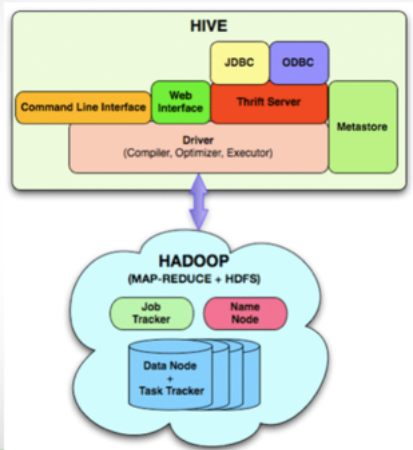
<https://data-flair.training/blogs/apache-hive-architecture/>

## 2. Hive Architecture

- 엄밀하게 보면, **Hive Web Interface**와 **CLI**는 오라클의 **SQL+** 같은 개념으로 클라이언트로 보는 경향도 있지만 제공해 주는 부분이기때문에 **Service**로 보자.
- **Hive Driver**는 **job**이 들어오면 실제 동작을 하고, 메타스토어의 구조를 파악한다.
- 메타스토어는 메타데이터를 저장하는 저장 공간이다.  
( 메타데이터 : 기본 데이터에 대한 정보 )
- 즉 실제 데이터가 저장되는 곳은 여기서는 **DerbyDB**이고 우리는 **mysql**에 저장한다.
- 그 저장된 데이터의 테이블 구조에 대한 정보 같은 메타데이터가 메타스토어에 저장되는 것이다.



## 2. Hive Architecture



## 2. Hive Architecture

### 3. 하이버 데이터 모델

- 하둡 상에 구축된 **정형화된 데이터를 관리하고** 쿼리하는 시스템
- 스토리지로 **HDFS에 저장**
- OLTP (online transaction proccession)에는 적합하지 않음

### 4. 데이터 관리

테이블 →→ HDFS의 디렉토리  
 파티션 →→ HDFS의 서브디렉토리  
 데이터 →→ HDFS의 파일

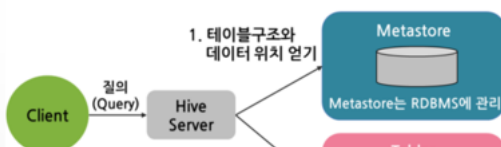
하이버테이블은 메타스토어(mariadb)에 테이블 구조인 스키마를 저장하고 데이터는 HDFS에 저장한다.



## 3. 하이버 데이터 처리 과정

### 1. 하이버 동작과정 - 1

- (1) 사용자의 **HiveQL** 명령어를 해석하여 맵리듀스 작업으로 변환
- (2) 메타스토어에서 테이블 구조와 데이터 위치를 얻음
- (3) 실제 데이터 질의 전달



2. 실제 데이터  
질의 전달files  
Data는 HDFS에 저장

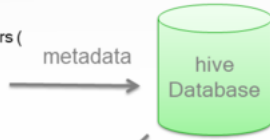
구자환 교수님 자료

### 3. 하이브 데이터 처리 과정

#### 1. 하이브 동작과정 - 2

- Hive는 HDFS의 data와 DB의 스키마를 Binding한다.

```
CREATE EXTERNAL TABLE users (
  user_id INT,
  age INT,
  gender STRING,
  occupation STRING,
  zip_code STRING
) ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
```



HDFS

```

1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43533
5|33|F|other|15213
  
```



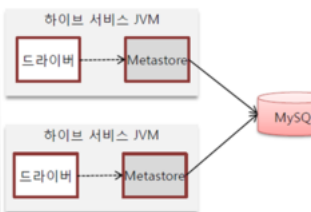
### 3. 하이브 데이터 처리 과정

#### 2. Hive Metastore

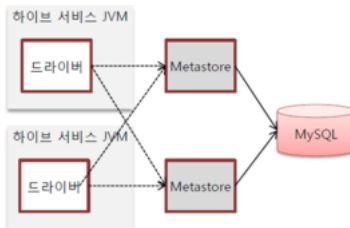


임베디드 메타스토어

- Hive는 별도 환경설정을 하지 않으면, Metastore는 Derby DB로 지정된다.
- Derby의 경우 파일 전체로 Lock이 되어 다수 이용자가 동시 사용 불가능하기 때문에 Metastore를 RDBMS로 지정한다.



로컬 메타스토어



원격 메타스토어

### 3. 하이브 데이터 처리 과정

#### 3. 하이브 기본 데이터 타입

데이터 타입	설명
<b>TINYINT</b>	1바이트 정수
<b>SMALLINT</b>	2바이트 정수
<b>INT</b>	4바이트 정수
<b>BIGINT</b>	8바이트 정수
<b>BOOLEAN</b>	TRUE/FALSE
<b>FLOAT</b>	소수점
<b>DOUBLE</b>	소수점
<b>STRING</b>	문자열

[참고] <http://hive.apache.org> > Language Manual



### 3. 하이브 데이터 처리 과정

#### 4. 하이브 Complex 데이터 타입

데이터 타입	설명
--------	----

<b>Array</b>	ARRAY<data_type>
<b>Map</b>	MAP<primitive_type, data_type>
<b>Struct</b>	STRUCT<col_name : data_type [COMMENT col_comment], ...>
<b>Union</b>	UNIONTYPE<data_type, data_type, ...> (Note: Only available starting with Hive 0.7.0.)



### 3. 하이브 데이터 처리 과정

#### 5. HiveQL vs SQL

(1) HiveQL에서는 **update**와 **delete** 안됨

데이터가 HDFS에 저장되는데 HDFS는 한번 저장한 파일을 수정 못함  
insert overwrite 를 사용하여 이미 입력된 데이터를 덮는 경우 가능

(2) HiveQL은 FROM 절에서만 서브 쿼리 사용가능

(3) HiveQL의 뷰는 읽기전용

(4) select 문에 having 안됨



#### [참고] Hive 와 Pig

Features	Hive	Pig
Language	SQL-like	PigLatin
Schemas/Types	Yes (explicit)	Yes (implicit)
Partitions	Yes	No
Server	Optional (Thrift)	No
User Defined Functions (UDF)	Yes (Java)	Yes (Java)
Custom Serializer/Deserializer	Yes	Yes
DFS Direct Access	Yes (implicit)	Yes (explicit)
Join/Order/Sort	Yes	Yes
Shell	Yes	Yes
Streaming	Yes	Yes
Web Interface	Yes	No
JDBC/ODBC	Yes (limited)	No

[출처 : <http://www.edureka.co/blog/pig-vs-hive/>]



### 4. HIVE 테이블 관리

#### 1. 내부테이블과 외부 테이블

##### 내부 테이블

` 하이브 데이터 웨어하우스에 저장됨  
( /hive/warehouse )

` 내부 테이블을 삭제하면 테이블의 메타정보와 테이블의 모든 데이터도 같이 삭제됨

` ORC 같은 최적화 형식으로 저장되어 비교적 성능 향상에 도움을 준다



## 4. HIVE 테이블 관리

### 외부 테이블

· 하이브가 직접 관리하지 않음

· 외부테이블의 데이터를 삭제하면 하이브의 테이블 메타 정의만 삭제되고 실제 데이터는 그대로 남는다.

· 테이블이 삭제되더라도 원본 데이터가 남아야 할 경우 주로 사용함

· 하이브 외에 다른 예코 시스템에서 사용하는 경우



## 4. HIVE 테이블 관리

### 2. 파티션과 버킷

#### 파티션

대용량 테이블을 논리적으로 나누어 효율적인 쿼리가 가능

#### 버킷

대용량의 데이터를 분할



## 4. HIVE 테이블 관리

### 3. 하이브 지원 파일 포맷

	텍스트파일	시퀀스파일	RC 파일	ORC 파일	파케이
저장기반	로우기반	로우기반	컬럼기반	컬럼기반	컬럼기반
압축	파일압축	레코드 / 블록압축	블록압축	블록압축	블록압축
스플릿지원	지원	지원	지원	지원	지원
압축적용시 스플릿지원	미지원	지원	지원	지원	지원
하이브 키워드	TEXTFILE	SEQUENCE FILE	RCFILE	ORCFILE	PARGUET

RC (record columnar )  
ORC (Optimized record-columnar)




## 4. HIVE 테이블 관리

### [ 참고 ] H catalog

- 하둡으로 생성한 데이터를 위한 테이블 및 스토리지 관리 서비스
- 하둡 예코 시스템들 간의 상호 운용성으르 높일 수 있도록 함
- 즉 메타스토어에 접근할 수 있는 역할

[참고] <http://hive.apache.org> > Language Manual







- HIVE 는 정형데이터를 처리한다

댓글

댓글알림

김용선

댓글을 남겨보세요



등록

---

자바쌤~!    <https://cafe.naver.com/javassem>