Cơ chế và nguyên lý Apache HIVE

Cách thức hoạt động Hive trong hadoop ecosystem

Hoạt động

Khi người dùng nhập ‘Query’ ở HIVE UI

Bước 1: các câu truy vấn được gửi đến Driver

Bước 2: Driver tạo ra 1 session cho truy vấn này và gửi query tới compiler để nhận execute plan (trong mỗi câu truy vấn trên hive đều phải có execution plan nó giống như là mã số để thực hiện câu lệnh vậy)

Bước 3: gửi yêu cầu lấy dữ liệu cần lấy của câu lệnh dưới metadata ở metastore

Bước 4: Metastore gửi lại dữ liệu cho compiler

Bước 5: Plan được sinh ra bởi compiler cần thiết để thực thi query sẽ được gửi lại tới phần thực thi (Execute engine)

Bước 6: Nhận lấy các execution plan từ driver và bắt đầu thực thi

Bước 6.1, 6.2, 6.3: xuống phần HDFS của hadoop ecosystem để thực thi quá trình thông qua mapreduceJob (đã được tìm hiểu qua bài hadoop mỗi câu lệnh khi muốn thực hiện ở mức HDFS phải có MapreduceJob)

Ở đây nó sẽ yêu cấu lấy các metadata cần thiết được phân tích ở bước 3, 4 và từ đó nó sẽ gọi Mapreduce thực thi công việc

Bước 7, 8, 9: Khi output được sinh ra, nó sẽ được ghi dưới dạng 1 temporary file, file này sẽ cung cấp các thông tin cần thiết cho các stages khác của plan

Mô hình dữ liệu trong hive

|  |  |
| --- | --- |
| Integer | Tinyint – 1 byte integer  Smallint – 2 byte integer  Int – 4 byte integer  Bigint – 8 byte integer |
| Boolean | True/false |
| Floating point number | Float – single precision  Double – double precision |
| String | sequence of characters in a specified character set |
| Others | structs  Maps  Arrays |

Tables

Trong hive thì có 2 loại tạo Table

1. Tạo table bình thường

Khi bạn tạo mới 1 tables thì Hive  sẽ chuyển các dữ liệu này tới tới kho dữ liệu của nó (warehouse directory)

CREATE TABLE managed\_table(dummy STRING);

LOAD DATA INPATH '/user/sample\_data/salary/salary.csv’INTO table managed\_tabl;

1. Tạo với external table

Với từ khóa EXTERNAL, Hive hiểu rằng nó không trực tiếp quản lý các dữ liệu này bởi vậy nó sẽ không thực hiện việc di chuyển chúng tới Hive’s data warehouse. Trên thực tế là Hive không thực hiện việc kiểm tra xem dữ liệu trên External localtion có tồn tại hay không. Đây là 1 tiện ích khác hữu dụng, nó cho phép ta có thể thực hiện việc khởi tạo dữ liệu khi  tạo 1 table trên  Hive. Và khi dữ liệu việc drop 1 external table thì Hive chỉ thực hiện việc xóa đi các metadata của nó.

create external table salary

(

salary\_id string,

employee\_id string,

payment double,

date string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' location '/user/sample\_data/salary/salary.csv'