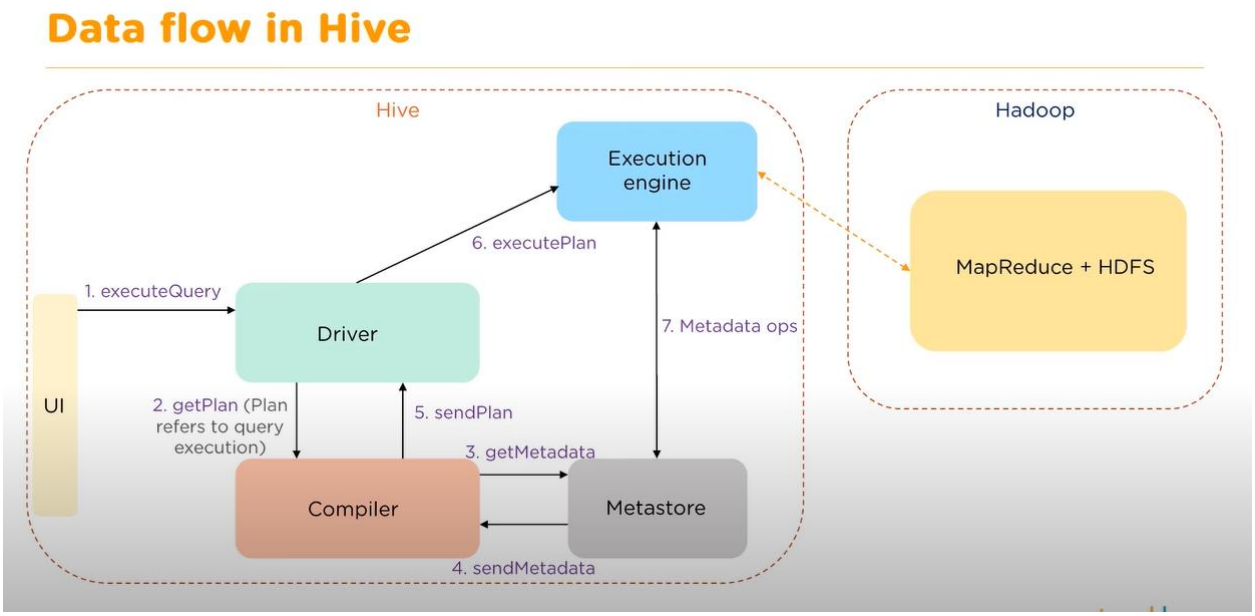
1. HOẠT ĐỘNG CỦA HIVE:

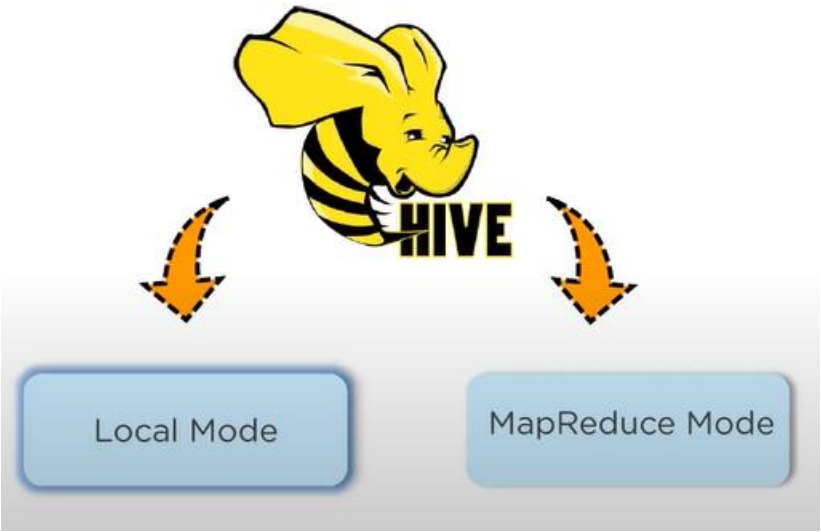


Hình 1. Mô hình hoạt động của Hive

Quy trình hoạt động của Hive có thể được mô tả theo các bước sau:

1. Các truy vấn tới từ User Interface (CLI, Hive Web Interface, ThirftServer) được gửi tới thành phần Driver (Bước 1 hình 3.1)
2. Driver tạo ra mới 1 session cho truy vấn này và gửi query tới compiler để nhận lấy Execution Plan (Bước 2 hình 3.1)
3. Compilter nhận các metadata cần thiết từ Metastore (Bước 3, 4 hình 3.1). Các metadata này sẽ được sử dụng để kiểm tra các biểu thức bên trong query mà Compiler nhận được.
4. Plan được sinh ra bởi Compiler (thông tin về các job (map-reduce) cần thiết để thực thi query sẽ được gửi lại tới thành phần thực thi (Bước 5hình 3.1)
5. Execution engine nhận yêu cầu thực thi và lấy các metadata cần thiết và yêu cầu mapreduce thực thi công việc (Bước 6.1, 6.2, 6.3 hình 3.1)
6. Khi output được sinh ra, nó sẽ được ghi dưới dạng 1 temporary file, temorary file này sẽ cung cấp các thông tin cần thiết cho các stages khác của plan. Nội dung của các temporary file này được execution đọc trực tiếp từ HDFS như là 1 phần của các lời gọi từ Driver (bước 7, 8, 9 hình 3.1)
7. Công cụ thực thi giao tiếp hai chiều với metastore để thực hiện các hoạt động như tạo, xóa bảng. Metastore lưu trữ thông tin về bảng, cột.

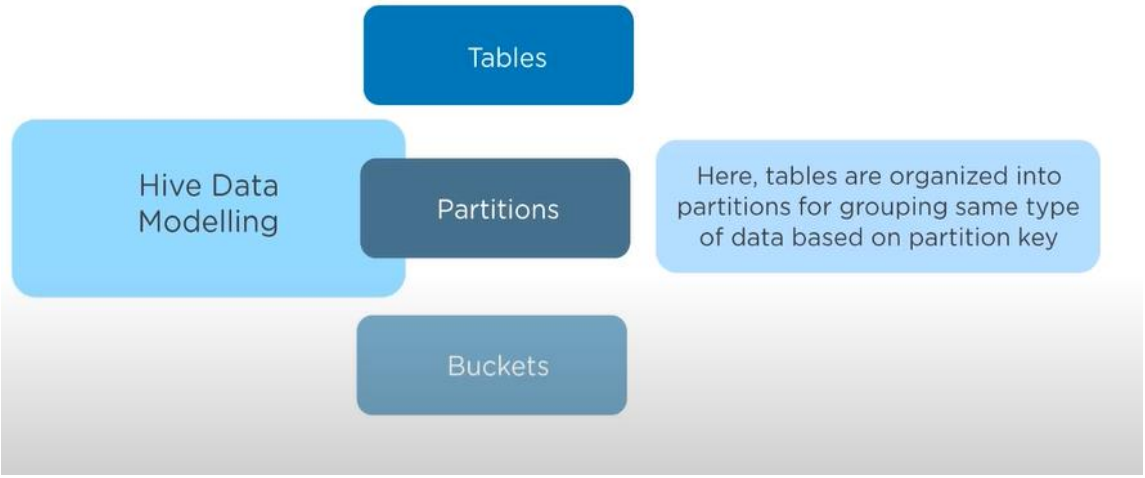
2. CÁC CHẾ ĐỘ HOẠT ĐỘNG TRONG HIVE.



Hình 2. Chế độ hoạt động

- Local Mode:
    - Được sử dụng khi hadoop có một nút dữ liệu và dữ liệu nhỏ.
    - Việc xử lý sẽ rất nhanh trên các bộ dữ liệu nhỏ hơn có trong máy cục bộ.
- MapReduce Mode:
    - Được sử dụng khi hadoop có nhiều nút dữ liệu và dữ liệu được trải trên các ghi chú dữ liệu khác nhau.
    - Xử lý bộ dữ liệu lớn có thể hiệu quả hơn khi sử dụng chế độ này.

3. MÔ HÌNH DỮ LIỆU TRONG HIVE:



Hình 3. Hive Data Model

Dữ liệu trong Hive được tổ chức thành các kiểu sau:

- Tables: tương tự như table trong các hệ cơ sở dữ liệu quan hệ. Trong Hive table có thể thực hiện các phép toán filter, join và union… Mặc định thì dữ liệu của Hive sẽ được lưu bên trong thư mục warehouse trên HDFS. Tuy nhiên Hive cũng cung cấp kiểu external table cho phép ta tạo ra và quản lý các table mà dữ liệu của nó đã tồn tại từ trước khi ta tạo ra table này hoặc nó được lưu trữ ở 1 thư mục khác bên trong hệ thống HDFS. Tổ chức row và column bên trong Hive có nhiều điểm tương đồng với tổ chức Row và Column trong các hệ cơ sở dữ liệu quan hệ. Hive có 2 kiểu table đó là: Managed Table và External tables.

- Partions: ở đây, các bảng được tổ chức thành các phần để nhóm cùng loại dữ liệu dựa trên khóa phân vùng.

  Ví dụ table web_log có thể phân chia dữ liệu của mình theo từng ngày là lưu dữ liệu của mỗi ngày trong 1 thư mục khác nhau bên dưới đường dẫn warehouse. Ví dụ: /warehouse/web_log/date=”01-01-2014″

- Buckets: Dữ liệu trong mỗi partion có thể được phân chia thành nhiều buckets khác nhau dựa trên 1 hash của 1 colume bên trong table. Mỗi bucket lưu trữ dữ liệu của nó bên dưới 1 thư mục riêng. Việc phân chia các partion thành các bucket giúp việc thực thi các query dễ dàng hơn.