Hive (http://hive.apache.org/)

Hive es un sistema de Data Warehouse para Hadoop que facilita el uso de la agregación de los datos, ad-hoc queries, y el análisis de grandes datasets almacenados en Hadoop. Hive proporciona métodos de consulta de los datos usando un lenguaje parecido al SQL, llamado HiveQL.

```
CREATE TABLE twitter_data(twitter_val STRING);

LOAD DATA INPATH "/input/data.txt" OVERWRITE INTO TABLE twitter_data;

SHOW TABLES;

DESCRIBE twitter_data;


SELECT * from twitter_data LIMIT 5;

select get_json_object(twitter_data.twitter_val, "$.created_at"),
get_json_object(twitter_data.twitter_val, "$.text") from twitter_data;



LOAD DATA INPATH '/home/info/drivers.csv' OVERWRITE INTO TABLE temp_drivers;

select * from temp_drivers limit 100;

CREATE TABLE drivers (driverId INT, name STRING, ssn BIGINT, location STRING,
certified STRING, wageplan STRING);

insert overwrite table drivers
SELECT
  regexp_extract(col_value, '^(?:([^,]*),?){1}', 1) driverId,
```

Ejemplo Hive

```
  regexp_extract(col_value, '^(?:([^,]*),?){2}', 1) name,
  regexp_extract(col_value, '^(?:([^,]*),?){3}', 1) ssn,
  regexp_extract(col_value, '^(?:([^,]*),?){4}', 1) location,
  regexp_extract(col_value, '^(?:([^,]*),?){5}', 1) certified,
  regexp_extract(col_value, '^(?:([^,]*),?){6}', 1) wageplan

from temp_drivers;

select * from temp_drivers limit 100;


CREATE TABLE temp_timesheet (col_value string);

LOAD DATA INPATH '/home/info/timesheet.csv' OVERWRITE INTO TABLE
temp_timesheet;
select * from temp_timesheet limit 100;

CREATE TABLE timesheet (driverId INT, week INT, hours_logged INT , miles_logged INT);

insert overwrite table timesheet
SELECT
  regexp_extract(col_value, '^(?:([^,]*),?){1}', 1) driverId,
  regexp_extract(col_value, '^(?:([^,]*),?){2}', 1) week,
  regexp_extract(col_value, '^(?:([^,]*),?){3}', 1) hours_logged,
  regexp_extract(col_value, '^(?:([^,]*),?){4}', 1) miles_logged

from temp_timesheet;

select * from temp_timesheet limit 100;

SELECT driverId, sum(hours_logged), sum(miles_logged) FROM timesheet GROUP BY
driverId;
```