10. Pig

You have a dataset containing daily weather data for different cities. Your task is to perform various queries to analyze the weather data.

Task 1: Create the data set in HDFS using Hue browser

```
New York,2023-01-01,25.0,60.0,0.1

New York,2023-01-02,28.0,65.0,0.0

New York,2023-01-03,30.0,70.0,0.2

Los Angeles,2023-01-01,20.0,55.0,0.0

Los Angeles,2023-01-02,22.0,50.0,0.1

Los Angeles,2023-01-03,25.0,45.0,0.0
```

Task 2: Load the weather data and define the schema.

```
weather_data = LOAD '/path/to/weather_data.csv' USING PigStorage(',') AS (

    city: chararray,

    date: chararray,

    temperature: double,

    humidity: double,

    precipitation: double

);
```

Task 3: Find the average temperature for each city

```
average_temperature_by_city = FOREACH (GROUP weather_data BY city) {

    generate group AS city, AVG(weather_data.temperature) AS average_temperature;

};
```

Task 4: Find the maximum temperature recorded for each city

```
max_temperature_by_city = FOREACH (GROUP weather_data BY city) {

    max_temp = MAX(weather_data.temperature);

    generate group AS city, max_temp AS max_temperature;

};
```

Task 5: Calculate the total precipitation for each month

```
total_precipitation_by_month = FOREACH (GROUP weather_data BY GetMonth(date)) {
    generate group AS month, SUM(weather_data.precipitation) AS total_precipitation;
};
```

Task 6: Find the city with the highest average humidity

```
max_humidity_city = ORDER weather_data BY humidity DESC;
top_humidity_city = LIMIT max_humidity_city 1;
```