Apache PIG: Group, Aggregate functions and Joins

In Apache Pig, group and aggregate functions are fundamental for data processing tasks, especially when dealing with large datasets. These functions allow you to group data based on certain criteria and perform calculations or operations on those grouped data.

Grouping (GROUP BY):

The GROUP BY operation in Pig is used to group data based on one or more fields in the dataset.

It collects all the tuples with the same key into a single group.

Aggregation:

After grouping the data, you often want to perform aggregate functions on each group to derive summary statistics or compute other aggregated values.

Pig provides several built-in aggregate functions like SUM, AVG, MIN, MAX, COUNT, etc.

Task 1: From the employers' data, find the sum of the salaries based upon each designation.

Create demo.txt file in cloudera

Id, fname, lname, field, salary

1,abc,xyz,manager,45120

Go to grunt (pig -x local)

A = LOAD 'path' USING PigStorage(',') AS (id:int, fname:chararray, lname:chararray, field:chararray, salary:int)

```
dump A;

b = group A by field;

dump b;

c = foreach b generate group, sum(A.salary);

dump c;
```

Task 2: Perform AVG, Max, Min and floor and ceil for the above task.

Task 3: Perform inner, left and right join with two tables.

Table1 metadata -> uid:int, fid:int, rating:int

Table2 metadata -> uid:int, name:chararray, city:chararray

After loading the tables into a and b variables, perform inner join as follows

```
c = join a by uid, b by uid;

dump c;

d = join a by uid left outer, b by uid;

dump d;
```