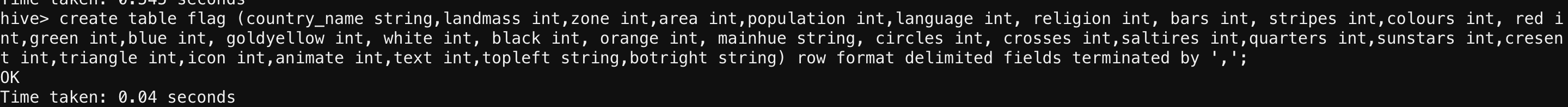
**Analysis of Flag Data using Hadoop-Hive and PIG**

1. **creating table youtube for flag.txt**:



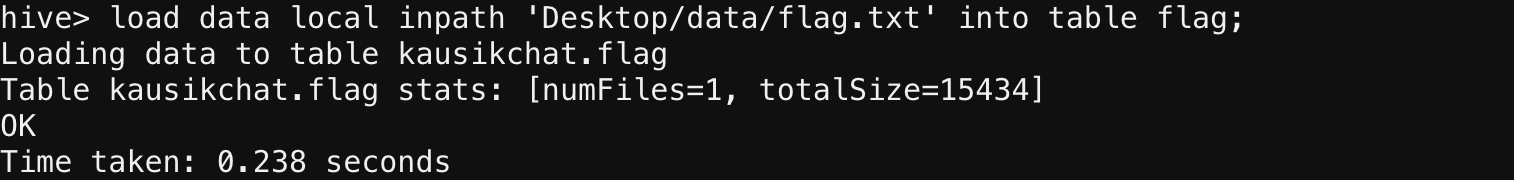
create table landmass (id int, land string) row format delimited fields terminated by ',';

create table zone (id int, name string) row format delimited fields terminated by ',';

create table language (id int, lan string) row format delimited fields terminated by ',';

create table religion (id int, rel string) row format delimited fields terminated by ',';

2. **loading data into youtube table**



insert into table landmass values (1,"N.America");

insert into table landmass values (2,"S.America");

insert into table landmass values (3,"Europe");

insert into table landmass values (4,"Africa");

insert into table landmass values (5,"Asia");

insert into table landmass values (6,"Oceania");

insert into table zone values (1,'NE'),(2,'SE'), (3,'SW'),(4,'NW');

insert into table language values (1,"English"),(2,"Spanish"),(3,"French"),(4,"German"),(5,"Slavic"),(6,"Other-Indo-European"),(7,"Chinese"),(8,"Arabic"),(9,"Japanese"),(10,"Others");

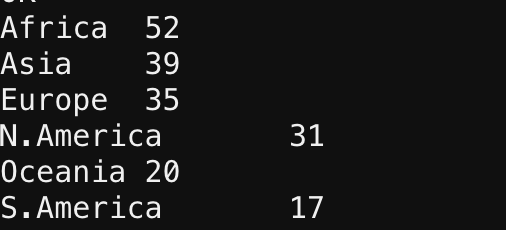
insert into table religion values (0,"Catholic"),(1,"Other Christian"),(2,"Muslim"),(3,"Budhist"),(4,"Hindu"),(5,"Ethnic"),(6,"Marxist"),(7,"Others");

**Problem solution**:

1. **Count number of countries based on landmass.**

**HIVE Solution:**

select b.land,count(a.country\_name) from flag a join landmass b on a.landmass = b.id group by b.land;



**PIG Solution:**

country\_data = LOAD 'data/flag.txt' USING PigStorage(',') as (name:chararray, landmass:int, zone:int, area:chararray, population:chararray, lang:int, rel:int, bar:int, stripes:int, col:int, red:int, green:int, blue:int, gold:int, white:int, black:int, orange:int, hue:chararray, circle:int, cross:int, saltire:int, quars:int, suns:int, cres:int, triangle:int, icon:int, animate:int, text:int, left:chararray, right:chararray);

grouped\_data = GROUP country\_data by landmass;

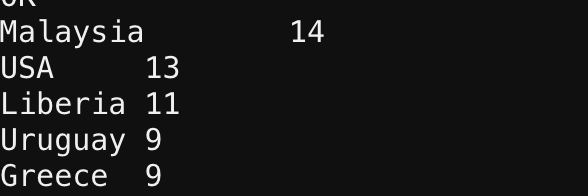
count\_data = FOREACH grouped\_data GENERATE group,COUNT(country\_data);

DUMP count\_data;

1. **Find out top 5 country with Sum of bars and strips in a flag.**

**HIVE Solution:**

select country\_name, SUM(bars+stripes) from flag group by country\_name order by country\_name limit 5;

****

PIG Solution:

foreach\_data = FOREACH country\_data GENERATE name,$7+$8;

grouped\_data = GROUP foreach\_data All;

bar\_data = FOREACH grouped\_data {

ord = ORDER foreach\_data BY $1 DESC;

top = LIMIT ord 5;

GENERATE FLATTEN(top);

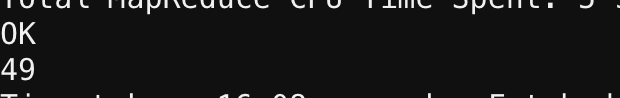
};

DUMP bar\_data;

**C. Count of countries with icon.**

**HIVE Solution:**

select count(\*) from flag where icon = 1;



**PIG Solution:**

filtered\_icon\_data = FILTER country\_data BY icon == 1;

grouped\_icon\_data = GROUP filtered\_icon\_data BY icon;

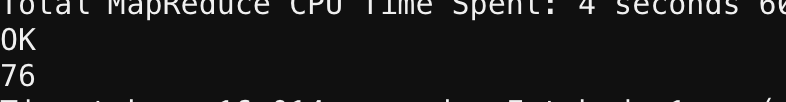
count\_icon\_data = FOREACH grouped\_icon\_data GENERATE group, COUNT(filtered\_icon\_data);

DUMP count\_icon\_data;

1. **Count of countries which have same top left and top right color in flag.**

**HIVE Solution:**

select count(\*) from flag where topleft = botright;



**PIG Solution:**

filtered\_flag\_data = FILTER country\_data BY $28 == $29;

grouped\_flag\_data = GROUP filtered\_flag\_data All;

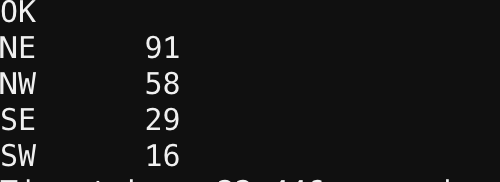
count\_flag\_data = FOREACH grouped\_flag\_data GENERATE group, COUNT(filtered\_flag\_data);

DUMP count\_flag\_data;

1. **Count number of countries based on zone.**

**HIVE Solution:**

select b.name,count(a.country\_name) from flag a join zone b on a.zone = b.id group by b.name;



**PIG Solution:**

grouped\_zone\_data = GROUP country\_data by landmass;

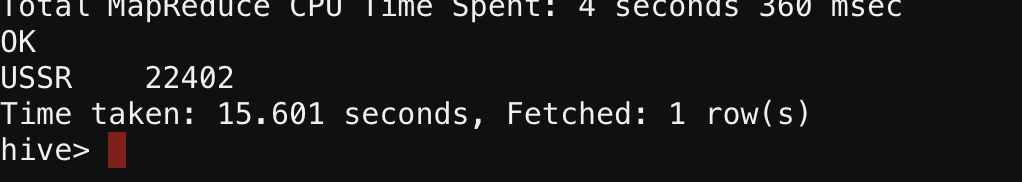
count\_zone\_data = FOREACH grouped\_zone\_data GENERATE group,COUNT(country\_data);

DUMP count\_zone\_data;

1. **Find out largest county in terms of area in NE zone.**

**HIVE Solution:**

select country\_name, area from flag where zone = 1 order by area desc limit 1;



**PIG Solution:**

filtered\_NE\_data = FILTER country\_data BY zone == 1;

grouped\_NE\_data = GROUP filtered\_NE\_data All;

max\_area\_data = FOREACH grouped\_NE\_data {

ord = ORDER filtered\_NE\_data BY $3 DESC;

top = LIMIT ord 1;

GENERATE FLATTEN(top);

};

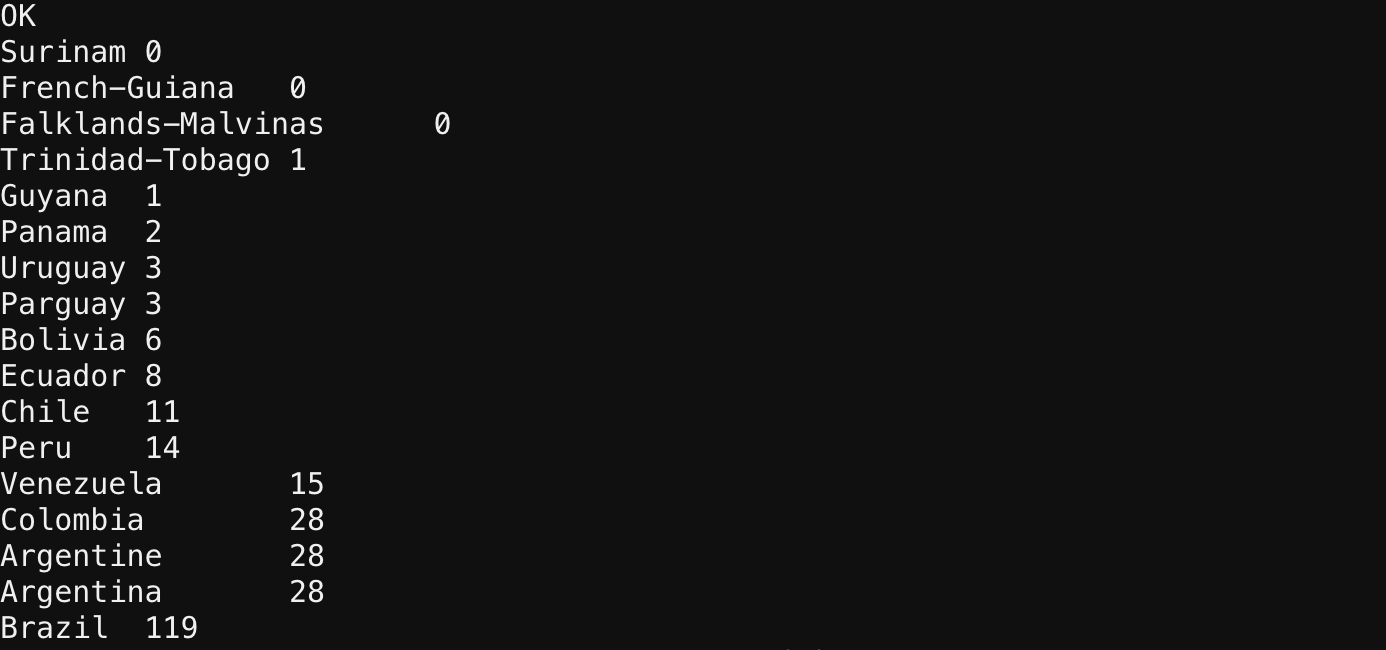
max\_area\_country = FOREACH max\_area\_data GENERATE name,area;

DUMP max\_area\_country;

1. **Find out least populated country in S.America landmass.**

**HIVE Solution:**

select country\_name, population from flag where landmass = 2 order by population asc;



**PIG Solution:**

filtered\_SA\_data = FILTER country\_data BY landmass == 2;

grouped\_SA\_data = GROUP filtered\_SA\_data All;

min\_population\_data = FOREACH grouped\_SA\_data {

ord = ORDER filtered\_SA\_data BY $4 ASC;

top = LIMIT ord 1;

GENERATE FLATTEN(top);

};

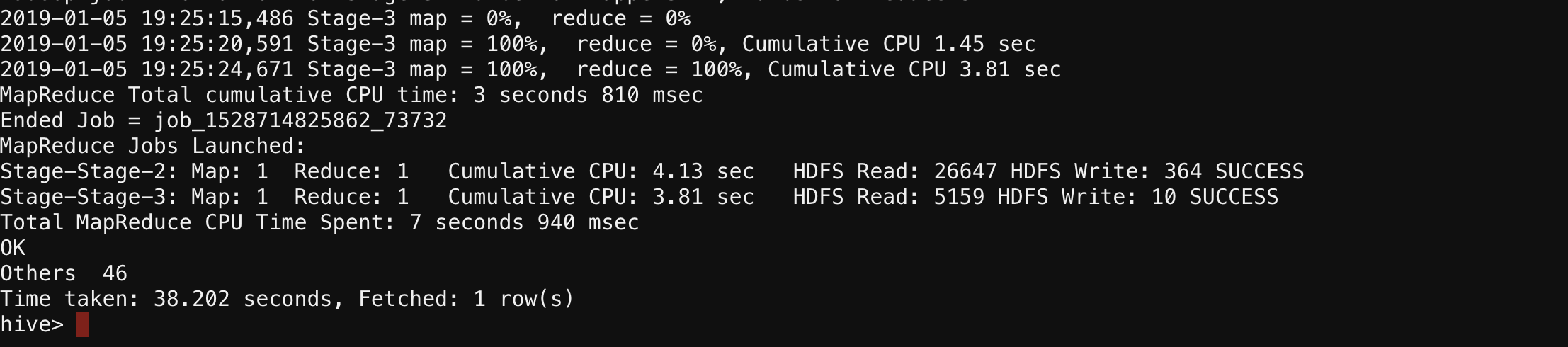
min\_population\_country = FOREACH min\_population\_data GENERATE name,population;

DUMP min\_population\_country;

1. **Find out largest speaking language among all countries.**

**HIVE Solution:**

select b.lan,count(\*) as cnt from flag a join language b on a.language = b.id group by b.lan order by cnt desc limit 1;



**PIG Solution:**

foreach\_data = FOREACH country\_data GENERATE name,lang;

grouped\_data = GROUP foreach\_data BY lang;

count\_data = FOREACH grouped\_data GENERATE group,(int)COUNT(foreach\_data);

grouped\_count\_data = GROUP count\_data All;

max\_lang\_data = FOREACH grouped\_count\_data {

ord = ORDER count\_data BY $1 DESC;

top = LIMIT ord 1;

GENERATE FLATTEN(top);

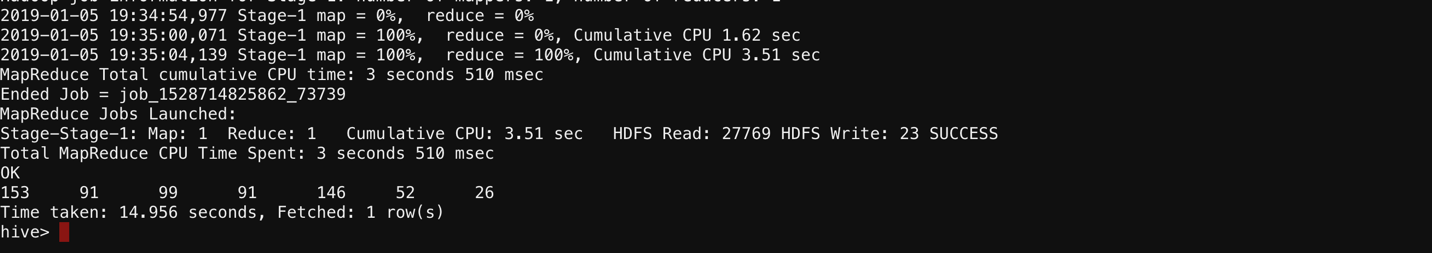
};

DUMP max\_lang\_data;

1. **Find most common color among flags from all countries.**

**HIVE Solution:**

Select sum(red) as red, sum(green) as green,sum(blue) as blue, sum(goldyellow),sum(white), sum(black), sum(orange) from flag;



**PIG Solution:**

grouped\_data = GROUP country\_data All;

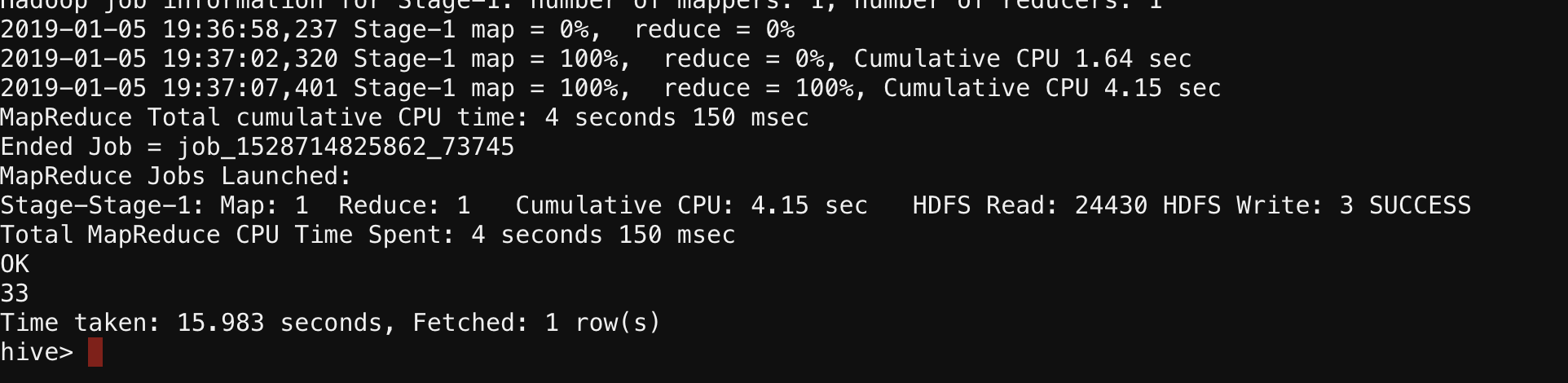
final\_color\_data = FOREACH grouped\_data GENERATE group,SUM(country\_data.$10),SUM(country\_data.$11),SUM(country\_data.$12),SUM(country\_data.$13),SUM(country\_data.$14),SUM(country\_data.$15),SUM(country\_data.$16);

dump final\_color\_data;

1. **Sum of all circles present in all country flags.**

**HIVE Solution:**

select sum(circles) from flag;



**PIG Solution:**

grouped\_circle\_data = GROUP country\_data All;

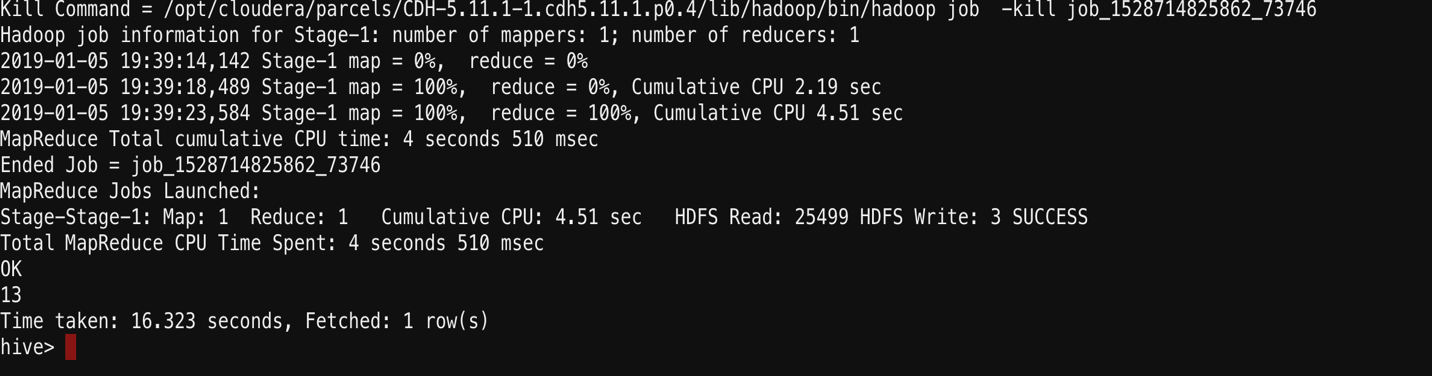
sum\_circle = FOREACH grouped\_circle\_data GENERATE group,SUM(country\_data.circle);

dump sum\_circle;

1. **Count of countries which have both icon and text in flag.**

**HIVE Solution:**

select count(country\_name) from flag where icon = 1 and text = 1;



**PIG Solution:**

filtered\_country\_data = FILTER country\_data BY ($25 == 1) AND ($29 == 1);

grouped\_text\_data = GROUP filtered\_country\_data All;

count\_text\_data = FOREACH grouped\_text\_data GENERATE group, COUNT(filtered\_country\_data);

DUMP count\_text\_data;