**Analyze monthly retail trade report for the US market**

A US‐based online retailer wants to launch a new product category and wants to understand the potential growth areas and areas that have stagnated over a period of time. It wants to use this information to ensure its product focus is aligned to opportunities that will grow over the next 5–7 years. The customer has also provided pointers to the data set you can use. The following are the goals for the project:

**‐ Analyze the entire data set and arrive at products that have experienced a consolidated yearly avg growth of 10% or more in sales since 2000**.

**‐ Analyze the entire data set and arrive at products that have experienced a consolidated yearly avg drop of 5% or more since 2000.**

**- Find top 5 products and bottom 5 products of overall sales for 3 years.**

Datasets are provided to you in text file format

Total there are 3 data sets for year 2000, 2001 and 2002

case\_study.pig

data\_2000 = load '/home/hduser/Downloads/retail/2000.txt' USING PigStorage(',') AS (sid,product,jan:double,feb:double,mar:double,apr:double,may:double,june:double,july:double,aug:double,sept:double,oct:double,nov:double,dec:double);

data\_2001 = load '/home/hduser/Downloads/retail/2001.txt' USING PigStorage(',') AS (sid,product,jan:double,feb:double,mar:double,apr:double,may:double,june:double,july:double,aug:double,sept:double,oct:double,nov:double,dec:double);

data\_2002 = load '/home/hduser/Downloads/retail/2002.txt' USING PigStorage(',') AS (sid,product,jan:double,feb:double,mar:double,apr:double,may:double,june:double,july:double,aug:double,sept:double,oct:double,nov:double,dec:double);

sales\_2000 = foreach data\_2000 generate $0,$1,($2+$3+$4+$5+$6+$7+$8+$9+$10+$11+$12+$13);

sales\_2001 = foreach data\_2001 generate $0,$1,($2+$3+$4+$5+$6+$7+$8+$9+$10+$11+$12+$13);

sales\_2002 = foreach data\_2002 generate $0,$1,($2+$3+$4+$5+$6+$7+$8+$9+$10+$11+$12+$13);

data\_join = join sales\_2000 by $0, sales\_2001 by $0, sales\_2002 by $0;

data\_join\_format = foreach data\_join generate $0,$1,$2,$5,$8;

group\_all = group data\_join\_format all;

total\_2000 = foreach group\_all generate SUM(data\_join\_format.$2) as total\_sum;

total\_2001 = foreach group\_all generate SUM(data\_join\_format.$3) as total\_sum;

total\_2002 = foreach group\_all generate SUM(data\_join\_format.$4) as total\_sum;

growth\_avg\_1 = foreach data\_join\_format generate $0,$1, ROUND\_TO((($3-$2)/$2)\*100,2);

growth\_avg\_2 = foreach data\_join\_format generate $0,$1, ROUND\_TO((($4-$3)/$3)\*100,2);

dump growth\_avg\_1;

dump growth\_avg\_2;

avg\_join = join growth\_avg\_1 by $0, growth\_avg\_2 by $0;

avg\_join\_format = foreach avg\_join generate $0,$1,$2,$5;

avg\_perc = foreach avg\_join\_format generate $0,$1,ROUND\_TO((($2+$3)/2),2);

filter\_product\_10 = filter avg\_perc by $2>10.0;

dump filter\_product\_10;

filter\_product\_0 = filter avg\_perc by $2<0;

filter\_product\_5 = filter filter\_product\_0 by $2<-5;

dump filter\_product\_5;

final\_sales\_format = foreach data\_join\_format generate $0,$1,($2+$3+$4);

order\_asc = order final\_sales\_format by $2 ASC;

order\_desc = order final\_sales\_format by $2 DESC;

topten = limit order\_desc 10;

bottomten = limit order\_asc 10;

dump topten;

dump bottomten;