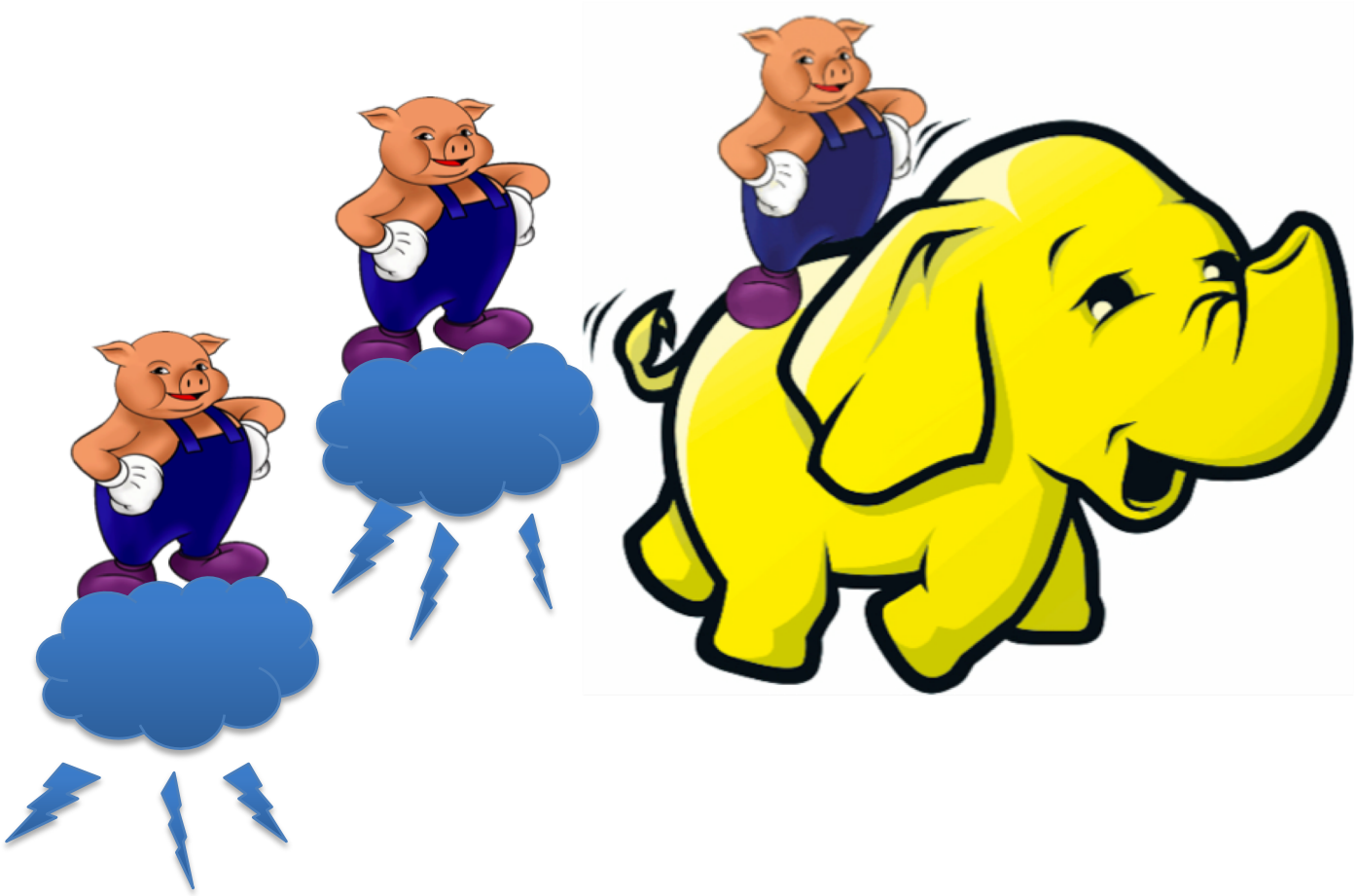
Pig Case Study



Some Baseball Data

We have two data files, Batting.csv (95195 lines of text) and Master.csv (17916 lines of text).

Batting.csv (partial)

```
aardsda01,2004,1,SFN,NL,11,11,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,11
aardsda01,2006,1,CHN,NL,45,43,2,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,45
aardsda01,2007,1,CHA,AL,25,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2
aardsda01,2008,1,BOS,AL,47,5,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,5
aardsda01,2009,1,SEA,AL,73,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
aardsda01,2010,1,SEA,AL,53,4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
aaronha01,1954,1,ML1,NL,122,122,468,58,131,27,6,13,69,2,2,28,39,,3,6,4,13,122
aaronha01,1955,1,ML1,NL,153,153,602,105,189,37,9,27,106,3,1,49,61,5,3,7,4,20,153
aaronha01,1956,1,ML1,NL,153,153,609,106,200,34,14,26,92,2,4,37,54,6,2,5,7,21,153
aaronha01,1957,1,ML1,NL,151,151,615,118,198,27,6,44,132,1,1,57,58,15,0,0,3,13,151
aaronha01,1958,1,ML1,NL,153,153,601,109,196,34,4,30,95,4,1,59,49,16,1,0,3,21,153
aaronha01,1959,1,ML1,NL,154,154,629,116,223,46,7,39,123,8,0,51,54,17,4,0,9,19,154
aaronha01,1960,1,ML1,NL,153,153,590,102,172,20,11,40,126,16,7,60,63,13,2,0,12,8,153
```

Master.csv (partial)

```
1,aaronha01,,aaronha01h,1934,2,5,USA,AL,Mobile,,,,,,,Hank,Aaron,,Henry Louis,"Hammer,Hammerin' Hank,Bad
Henry",180,72,R,R,4/13/1954,10/3/1976,,aaronha01,aaronha01,aaroh101,aaronha01,aaronha01
2,aaronto01,,,1939,8,5,USA,AL,Mobile,1984,8,16,USA,GA,Atlanta,Tommie,Aaron,,Tommie Lee,,190,75,R,R,
4/10/1962,9/26/1971,,aaronto01,aaronto01,aarot101,aaronto01,aaronto01
3,aasedo01,,,1954,9,8,USA,CA,Orange,,,,,,,Don,Aase,,Donald William,,190,75,R,R,7/26/1977,10/3/1990,Cal St.
Fullerton,aasedo01,aasedo01,aased001,aasedo01,aasedo01
4,abadan01,,,1972,8,25,USA,FL,West Palm Beach,,,,,,,Andy,Abad,,,,184,73,L,L,9/10/2001,4/13/2006,Middle
Georgia JC,abadan01,abadan01,abada001,abadan01,abadan01
5,abadijo01,,,1854,11,4,USA,PA,Philadelphia,1905,5,17,USA,NJ,Pemberton,John,Abadie,,John,,192,72,R,R,
4/26/1875,6/10/1875,,abadijo01,abadijo01,abadj101,abadijo01,abadijo01
6,abbated01,,,1877,4,15,USA,PA,Latrobe,1957,1,6,USA,FL,Ft.Lauderdale,Ed,Abbaticchio,,Edward James,Batty,
170,71,R,R,9/4/1897,9/15/1910,,abbated01,abbated01,abbae101,abbated01,abbated01
```

Analysis Goals

- Find the player with the highest run for each year.
- What are the First and last names of the player who had the highest run in each year? Display the name, player id, year and the highest run.
- Who (First name and last name) and in what year had the highest run of all the years in the dataset?
- What is the average run (rounded to two decimal points) of each year?
- What is the average run (rounded to two decimal points) of all the years in the dataset?

Load the Data and Create the Relations

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'
    using PigStorage(',')
    AS (player_id:chararray, year:int,
    c3:int,c4:chararray,c5:chararray,
    c6:int,c7:int,c8:int,run:int,c10:int,
    c11:int,c12:int,c13:int,c14:int,
    c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);

B = foreach A generate player_id, year, run;
DUMP B; --check the result
```

The Beginning Portion of Relation B

```
(aardsda01,2004,0)
(aardsda01,2006,0)
(aardsda01,2007,0)
(aardsda01,2008,0)
(aardsda01,2009,0)
(aardsda01,2010,0)
(aaronha01,1954,58)
(aaronha01,1955,105)
(aaronha01,1956,106)
(aaronha01,1957,118)
(aaronha01,1958,109)
(aaronha01,1959,116)
(aaronha01,1960,102)
(aaronha01,1961,115)
(aaronha01,1962,127)
(aaronha01,1963,121)
```

This Achieved the Same Result as in Hive Case Study

	default.batting.player_id	default.batting.year	default.batting.runs
0	aardsda01	2004	0
1	aardsda01	2006	0
2	aardsda01	2007	0
3	aardsda01	2008	0
4	aardsda01	2009	0
5	aardsda01	2010	0
6	aaronha01	1954	58
7	aaronha01	1955	105
8	aaronha01	1956	106
9	aaronha01	1957	118
10	aaronha01	1958	109
11	aaronha01	1959	116
12	aaronha01	1960	102
13	aaronha01	1961	115

Another Way to Create the Relations

```
1 A = load '/mydata/lahman591-csv/Batting.csv'
2     using PigStorage(',');
3 B = FOREACH A GENERATE $0 as
4     playerID, $1 as year, $8 as runs;
5 dump B;
```

The Beginning Portion of Relation B

```
(aardsda01,2004,0)
(aardsda01,2006,0)
(aardsda01,2007,0)
(aardsda01,2008,0)
(aardsda01,2009,0)
(aardsda01,2010,0)
(aaronha01,1954,58)
(aaronha01,1955,105)
(aaronha01,1956,106)
(aaronha01,1957,118)
(aaronha01,1958,109)
(aaronha01,1959,116)
(aaronha01,1960,102)
(aaronha01,1961,115)
(aaronha01,1962,127)
(aaronha01,1963,121)
```

Find the Highest Run for Each Year

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run);
dump D;
```

The Beginning Portion of Relation D, MAX(run) by Year

```
(1871,66)
(1872,94)
(1873,125)
(1874,91)
(1875,115)
(1876,126)
(1877,68)
(1878,60)
(1879,85)
(1880,91)
(1881,86)
(1882,99)
(1883,110)
(1884,160)
(1885,130)
(1886,155)
(1887,167)
(1888,134)
(1889,152)
```

Find The Highest Run for Each Year with Player ID

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run) as maxRun;
E = join B by (year,run), D by (group, maxRun);


dump E;
```

Find The Highest Run for Each Year with Player ID (Beginning)

```
(barnero01,1871,66,1871,66)
(eggleda01,1872,94,1872,94)
(barnero01,1873,125,1873,125)
(mcveyca01,1874,91,1874,91)
(barnero01,1875,115,1875,115)
(barnero01,1876,126,1876,126)
(orourji01,1877,68,1877,68)
(highadi01,1878,60,1878,60)
(jonesch01,1879,85,1879,85)
(dalryab01,1880,91,1880,91)
(gorege01,1881,86,1881,86)
(gorege01,1882,99,1882,99)
(stoveha01,1883,110,1883,110)
(dunlafr01,1884,160,1884,160)
(stoveha01,1885,130,1885,130)
(kellyki01,1886,155,1886,155)
(oneilti01,1887,167,1887,167)
(pinknge01,1888,134,1888,134)
(stoveha01,1889,152,1889,152)
(griffmi01,1889,152,1889,152)
```

Find The Highest Run for Each Year with Player ID (End)

```
(sosasa01,1998,134,1998,134)
(bagweje01,1999,143,1999,143)
(bagweje01,2000,152,2000,152)
(sosasa01,2001,146,2001,146)
(soriaal01,2002,128,2002,128)
(pujolal01,2003,137,2003,137)
(pujolal01,2004,133,2004,133)
(pujolal01,2005,129,2005,129)
(sizemgr01,2006,134,2006,134)
(rodrial01,2007,143,2007,143)
(ramirha01,2008,125,2008,125)
(pujolal01,2009,124,2009,124)
(pujolal01,2010,115,2010,115)
(grandcu01,2011,136,2011,136)
```

Just Show the Three Columns: Player_id, Year and MaxRun

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run) as maxRun;
E = join B by (year,run), D by (group, maxRun);
F = foreach E generate $0, $1, $2;

dump F;
```

Just Show the Three Columns: Player_id, Year and MaxRun

```
(barnero01,1871,66)
(eggleda01,1872,94)
(barnero01,1873,125)
(mcveyca01,1874,91)
(barnero01,1875,115)
(barnero01,1876,126)
(orourji01,1877,68)
(highadi01,1878,60)
(jonesch01,1879,85)
(dalryab01,1880,91)
(gorege01,1881,86)
(gorege01,1882,99)
(stoveha01,1883,110)
(dunlafr01,1884,160)
(stoveha01,1885,130)
(kellyki01,1886,155)
(oneilti01,1887,167)
(pinknge01,1888,134)
(stoveha01,1889,152)
```

Load Data from Master.csv

```
G = LOAD '/mydata/lahman591-csv/Master.csv'
    USING PigStorage(',')
    as (lahmanID:int,playerID:chararray,managerID:
    chararray,hofID:chararray,birthYear:int,
    birthMonth:int,birthDay:int,birthCountry:chararray,
    birthState:chararray,birthCity:chararray,
    deathYear:int,deathMonth:int,deathDay:int,
    deathCountry:chararray,deathState:chararray,
    deathCity:chararray,nameFirst:chararray,
    nameLast:chararray,nameNote:chararray,
    nameGiven:chararray,nameNick:chararray,weight:float,
    height:float,bats:chararray,throws:chararray,
    debut:chararray,finalGame:chararray,
    college:chararray,lahman40ID:chararray,
    lahman45ID:chararray,retroID:chararray,
    holtzID:chararray,bbrefID:chararray);
dump G;
```

A Relation Was Created Successfully

```
(1,aaronha01,,aaronha01h,1934,2,5,USA,AL,Mobile,,,,,,,Hank,Aaron,,Henry Louis,HammerHammerin'
 HankBad Henry,180,72,R,R,4/13/1954,10/3/1976,,aaronha01,aaronha01,aaroh101,aaronha01,aaronha
01)
(2,aaronto01,,,1939,8,5,USA,AL,Mobile,1984,8,16,USA,GA,Atlanta,Tommie,Aaron,,Tommie Lee,,190,
75,R,R,4/10/1962,9/26/1971,,aaronto01,aaronto01,aarot101,aaronto01,aaronto01)
(3,aasedo01,,,1954,9,8,USA,CA,Orange,,,,,,,Don,Aase,,Donald William,,190,75,R,R,7/26/1977,10/
3/1990,Cal St. Fullerton,aasedo01,aasedo01,aased001,aasedo01,aasedo01)
(4,abadan01,,,1972,8,25,USA,FL,West Palm Beach,,,,,,,Andy,Abad,,,,184,73,L,L,9/10/2001,4/13/2
006,Middle Georgia JC,abadan01,abadan01,abada001,abadan01,abadan01)
(5,abadijo01,,,1854,11,4,USA,PA,Philadelphia,1905,5,17,USA,NJ,Pemberton,John,Abadie,,John,,19
2,72,R,R,4/26/1875,6/10/1875,,abadijo01,abadijo01,abadj101,abadijo01,abadijo01)
(6,abbated01,,,1877,4,15,USA,PA,Latrobe,1957,1,6,USA,FL,Ft.Lauderdale,Ed,Abbaticchio,,Edward
James,Batty,170,71,R,R,9/4/1897,9/15/1910,,abbated01,abbated01,abbae101,abbated01,abbated01)
(7,abbeybe01,,,1869,11,29,USA,VT,Essex,1962,6,11,USA,VT,Essex Junction,Bert,Abbey,,Bert Wood,
,175,71,R,R,6/14/1892,9/23/1896,,abbeybe01,abbeybe01,abbeb101,abbeybe01,abbeybe01)
(8,abbeych01,,,1866,10,14,USA,NE,Falls City,1926,4,27,USA,CA,San Francisco,Charlie,Abbey,,Cha
rles S.,,169,68,L,,8/16/1893,8/19/1897,,abbeych01,abbeych01,abbec101,abbeych01,abbeych01)
(9,abbotda01,,,1862,3,16,USA,OH,Portage,1930,2,13,USA,MI,Ottawa Lake,Dan,Abbott,,Leander Fran
klin,Big Dan,190,71,R,R,4/19/1890,5/23/1890,,abbotda01,abbotda01,abbod101,abbotda01,abbotda01
)
(10,abbotfr01,,,1874,10,22,USA,OH,Versailles,1935,6,11,USA,CA,Los Angeles,Fred,Abbott,born Ha
rry Frederick Winbigler,Harry Frederick,180,70,R,R,4/25/1903,9/20/1905,abbotfr01,abbotfr01
```

The 2nd Goal

- What are the First and last names of the player who had the highest run in each year? Display the name, player id, year and the highest run.

The Pig Script

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run) as maxRun;
E = join B by (year,run), D by (group, maxRun);
--create a relation with player id, year, and max run
F = foreach E generate $0, $1, $2;




H = foreach G generate playerid, namefirst, namelast;
I = join H by playerid, F by player_id;
--create a relation first name, last name, player id, year, and max run
J = foreach I generate $1, $2, $3, $4, $5;
dump J;
```

The Result

```
(Ross,Barnes,barnero01,1871,66)
(Dave,Eggler,eggleda01,1872,94)
(Ross,Barnes,barnero01,1873,125)
(Cal,McVey,mcveyca01,1874,91)
(Ross,Barnes,barnero01,1875,115)
(Ross,Barnes,barnero01,1876,126)
(Jim,O'Rourke,orourji01,1877,68)
(Dick,Higham,highadi01,1878,60)
(Charley,Jones,jonesch01,1879,85)
(Abner,Dalrymple,dalryab01,1880,91)
(George,Gore,gorege01,1881,86)
(George,Gore,gorege01,1882,99)
(Harry,Stovey,stoveha01,1883,110)
(Fred,Dunlap,dunlafr01,1884,160)
(Harry,Stovey,stoveha01,1885,130)
(King,Kelly,kellyki01,1886,155)
(Tip,O'Neill,oneilti01,1887,167)
(George,Pinkney,pinknge01,1888,134)
(Mike,Griffin,griffmi01,1889,152)
```

The 3rd Goal

- Who (First name and last name) and what year had the highest run of all the years in the dataset?

The Pig Script to Find the All-Year Max Run

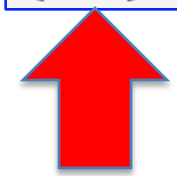
```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run) as maxRun;
E = join B by (year,run), D by (group, maxRun);
--create a relation with player id, year, and max run
F = foreach E generate $0, $1, $2;



H = foreach G generate playerid, namefirst, namelast;
I = join H by playerid, F by player_id;
--create a relation first name, last name, player id, year, and max run
J = foreach I generate $1, $2, $3, $4, $5;
K = order J by year;
describe K;
--find out the highest run of all the years
L = group K all;
describe L;
M = foreach L generate MAX(K.run);
dump M;
```

The All Year Highest Run

```
K: {H::namefirst: chararray,H::namelast: chararray,F::B::player
_id: chararray,F::B::year: int,F::B::run: int}
L: {group: chararray,K: {(H::namefirst: chararray,H::namelast:
chararray,F::B::player_id: chararray,F::B::year: int,F::B::run:
 int)}}
(192)
```



Pig Script to Locate the Player Name and Year of the All-Year Max Run

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
B = foreach A generate player_id, year, run;
C = group B by year;
D = foreach C generate group, MAX(B.run) as maxRun;
E = join B by (year,run), D by (group, maxRun);
--create a relation with player id, year, and max run
F = foreach E generate $0, $1, $2;



H = foreach G generate playerid, namefirst, namelast;
I = join H by playerid, F by player_id;
--create a relation first name, last name, player id, year, and max run
J = foreach I generate $1, $2, $3, $4, $5;
K = order J by year;
--find out the highest run of all the years
L = group K all;
M = foreach L generate MAX(K.run) as allMaxRun;
N = join J by run, M by allMaxRun;
O = foreach N generate $0, $1, $3, $4;
dump O;
```

The Result

```
(Billy,Hamilton,1894,192)
```

The 4th Goal

- What is the average run (rounded to two decimal points) of each year?

The Pig Script to Find Year Average Run

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
C = group A by year;
D = foreach C generate group, ROUND_TO(AVG(A.run), 2);
dump D;
```

The Year Average Run

```
(1871,23.12)
(1872,21.73)
(1873,28.64)
(1874,28.21)
(1875,19.42)
(1876,24.73)
(1877,21.03)
(1878,23.8)
(1879,26.84)
(1880,23.64)
(1881,25.95)
(1882,24.28)
(1883,31.91)
(1884,21.41)
(1885,26.32)
(1886,32.8)
(1887,40.28)
(1888,30.28)
(1889,37.97)
```

The 5th Goal

- What is the average run (rounded to two decimal points) of all the years in the dataset?

The Script to Calculate All-Year Average Run

```
A = LOAD '/mydata/lahman591-csv/Batting.csv'using PigStorage(',')
    AS (player_id:chararray, year:int, c3:int,c4:chararray,
    c5:chararray,c6:int,c7:int,c8:int,run:int,c10:int,c11:int,
    c12:int,c13:int,c14:int,c15:int,c16:int,c17:int,c18:int,c19:int,
    c20:int,c21:int,c22:int,c23:int,c24:int);
C = group A all;
D = foreach C generate group, ROUND_TO(AVG(A.run), 2);
dump D;
```

The All-Year Average Run

```
(all,20.59)
```