Indian Premier League T20

The IPL dataset contains cricket match played between Gujarat Lions Vs Sunrisers Hyderabad on 27/5/16 in Delhi.

Understanding Data

The data format is comma separated values.
The top 19 rows contain data on the other details like participating team, venue, city and results of the match.

It contains 12 columns made up of following: Ball, team batting, ball overs, bating team, batsmen on strike, Second batsmen, bowler, runs scored, extras, result, player who is out.

Analysing data with Apache Pig

1) LOAD the IPL data set.
   grunt> load_ipldata = LOAD
   '/home/cloudera/khasimbabu/PIG_Excercise/CaseStudy5/attachment_Dataset-Case-Study-Apache-Pig.xls' USING PigStorage(',') as (ball:chararray, team:int, overs: float, teamBating:chararray,batsmen1:chararray, batsmen2:chararray, bowler:chararray,runs:int,other:int, result: chararray, playerOut:chararray);

2) How many runs did David Warner score in the match?
   grunt> filterDavid = FILTER load_ipldata BY batsmen1 MATCHES '^DA.+';
   grunt> runsByDavid = FOREACH (GROUP filterDavid by 'DA Warner')GENERATE
   SUM(filterDavid.runs) as TotalRuns;
   grunt> describe runsByDavid;
   runsByDavid: {TotalRuns: long}

3) How many wickets did DJ Bravo take in match?
   grunt> ipl_bravodetails = FILTER load_ipldata by bowler=='DJ Bravo' AND result!='runout';
   grunt> ipl_bravowickets = FOREACH (GROUP ipl_bravodetails ALL) GENERATE
   COUNT(ipl_bravodetails);
   grunt> dump ipl_bravowickets;

4) How many runs did Praveen Kumar concede in the match?
   grunt> PraveenRunsFilter = FILTER load_ipldata by bowler=='P Kumar' AND runs!=0;
   grunt> count_PraveenRuns = FOREACH (GROUP PraveenRunsFilter ALL) GENERATE
   SUM(PraveenRunsFilter.runs);
   grunt> dump count_PraveenRuns;

5) How many extras were bowled in the match?
   grunt> ipl_extrabowled = FOREACH (GROUP ipl_extrasbowled_filter ALL) GENERATE
   COUNT(ipl_extrasbowled_filter);

6) What is the final score of Gujarat Lions?
   grunt> lionsteam_filter = FILTER load_ipldata by teamBating=='Gujarat Lions';

```
grunt> lionsteam_runs = FOREACH (GROUP lionsteam_filter ALL) GENERATE
SUM(lionsteam_filter.runs);
```

7) How many balls were faced by Bb McCullum in the match?
```
grunt> ballsfaced_McCul_filter = FILTER load_ipldata by batsmen1=='BB McCullum';
grunt> grp = GROUP ballsfaced_McCul_filter by 'BB McCullum'
grunt> ballsfaced_McCul = FOREACH grp GENERATE
COUNT(ballsfaced_McCul_filter.batsmen1);
grunt> dump ballsfaced_McCul;
```

8) Find the players who were run out in the match
```
grunt> runout_player_filter = FILTER load_ipldata by result=='run out';
grunt> dump runout_player_filter;
```

9) Store the players who were run out in the match in HDFS
```
grunt> STORE runout_player_filter into
'/home/cloudera/khasimbabu/PIG_Excercise/ipl_runout' USING PigStorage(',');
```