

DA Assignment1

Group member: Chang Tian(cht97), Huan Chen(huc48), Jiayu Chen(jic117)

1. Describe the schemas of your relations (CREATE statements)

```
CREATE TABLE `players` (  
  `playerid` varchar(45) NOT NULL,  
  `namefirst` varchar(45) DEFAULT NULL,  
  `namelast` varchar(45) DEFAULT NULL,  
  PRIMARY KEY (`playerid`)  
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

```
CREATE TABLE `batting` (  
  `playerid` varchar(45) DEFAULT NULL,  
  `yearid` int(11) DEFAULT NULL,  
  `teamid` varchar(45) DEFAULT NULL,  
  `games` int(11) DEFAULT NULL,  
  `atbats` int(11) DEFAULT NULL,  
  `runsscored` int(11) DEFAULT NULL,  
  `hits` int(11) DEFAULT NULL,  
  `homeruns` int(11) DEFAULT NULL,  
  `runsbattedin` int(11) DEFAULT NULL,  
  `walks` int(11) DEFAULT NULL,  
  `strikeouts` int(11) DEFAULT NULL,  
  `battingid` int(11) NOT NULL AUTO_INCREMENT,  
  PRIMARY KEY (`battingid`)  
) ENGINE=InnoDB AUTO_INCREMENT=96184 DEFAULT CHARSET=latin1;
```

```
CREATE TABLE `halloffame` (  
  `playerid` varchar(45) DEFAULT NULL,  
  `yearid` int(11) DEFAULT NULL,  
  `votedby` varchar(45) DEFAULT NULL,  
  `ballots` int(11) DEFAULT NULL,  
  `needed` int(11) DEFAULT NULL,  
  `votes` int(11) DEFAULT NULL,  
  `inducted` varchar(45) DEFAULT NULL,  
  `category` varchar(45) DEFAULT NULL,  
  `halloffameid` int(11) NOT NULL AUTO_INCREMENT,  
  PRIMARY KEY (`halloffameid`)  
) ENGINE=InnoDB AUTO_INCREMENT=4121 DEFAULT CHARSET=latin1;
```

2. Explain your data import strategy (python library/csv importer/shell command/etc)

Using MySQL WorkBench to import csv files as tables in schema.

3. Queries / answers to questions (answer first query after)

1. a) how many players are inducted into the hall of fame?

count(distinct playerid)
▶ 312

```
select count(distinct playerid)  
from halloffame  
where inducted = 'Y';
```

b) how many distinct players in the batting data set are in the hall of fame?

count(temp.playerid)
▶ 248

```
select count(temp.playerid)
from (select distinct(h.playerID) as playerid
      from halloffame h join batting b on h.playerid=b.playerid
      where h.inducted='y') as temp;
```

2. a) what player has exactly 3000 hits total for his career?

playerid
► clemere01

```
select playerid
from (select playerid, sum(hits) hit
      from batting
      group by playerid) as temp
where temp.hit = 3000;
```

b) what year was he inducted into the hall of fame?

yearid
► 1973

```
select h.yearid
from halloffame h
where h.playerid = (select playerid
                   from (select playerid, sum(hits) hit
                         from batting
                         group by playerid) as temp
                   where temp.hit = 3000);
```

3. what player had the highest number of homeruns in 1919?

playerID	total_homeruns
► ruthba01	29

```
select temp.playerID, temp.total_homeruns
from (select sum(b.Homeruns) as total_homeruns, b.playerID
      from Batting b
      where b.yearID=1919
      group by b.playerID) as temp
where temp.total_homeruns in (select max(temp2.total_homeruns) from (select sum(b.Homeruns) as total_homeruns, b.playerID
                             from Batting b
                             where b.yearID=1919
                             group by b.playerID ) as temp2);
```

4. what team has had the most hits after the year 2000, how many?

teamid	total_hits
► BOS	23029

```
select b.teamid, sum(b.hits) total_hits from batting b
where b.yearid > 2000
group by b.teamid
order by total_hits desc
limit 1;
```

5. a) what player has the most cumulative hits?

playerid
► rosepe01

```

select temp.playerid
from (select sum(b.hits) as summation, b.playerid
      from batting b
      group by b.playerid) as temp
where temp.summation in (select max(temp2.summation)
                        from (select sum(b.hits) as summation, b.playerid
                              from batting b group by b.playerid) as temp2);

```

b) how many does he have?

max(temp2.summation)
▶ 4256

```

select max(temp2.summation)
from (select sum(b.hits) as summation, b.playerid
      from batting b group by b.playerid) as temp2;

```

c) is he in the hall of fame?

playerid	inducted
▶ rosepe01	N

```

select distinct h.playerid, h.inducted
from halloffame h
where h.playerid = (select temp.playerid
                   from (select sum(b.hits) as summation, b.playerid
                         from batting b
                         group by b.playerid) as temp
                   where temp.summation in (select max(temp2.summation)
                                             from (select sum(b.hits) as summation, b.playerid
                                                   from batting b
                                                   group by b.playerid) as temp2));

```

d) how many votes has he received total?

sum(h.votes)
▶ 74

```

select sum(h.votes)
from halloffame h
where h.playerid in (select temp.playerid
                    from (select sum(b.hits) as summation, b.playerid
                          from batting b
                          group by b.playerid) as temp
                    where temp.summation in (select max(temp2.summation)
                                              from (select sum(b.hits) as summation, b.playerid
                                                    from batting b
                                                    group by b.playerid) as temp2));

```

6. a) what 3 players have the highest number of cumulative strikeouts over their career that have been inducted into the hall of fame?

playerid
▶ jacksre01
stargwi01
schmimi01

```

select B.playerid
from batting B, halloffame H
where inducted='Y' and H.playerid=B.playerid
group by playerid
order by sum(strikeouts) desc
limit 3;

```

b) how many total strikeouts do they each have?

playerid	total_strikeouts
jacksre01	2597
stargwi01	1936
schmimi01	1883

```
select B.playerid, sum(strikeouts) as total_strikeouts
from batting B, halloffame H
where inducted='Y' and H.playerid=B.playerid
group by playerid
order by total_strikeouts desc
limit 3;
```

7. a) what 3 players have the highest cumulative total number of homeruns and have been inducted into the hall of fame?

playerid
aaronha01
ruthba01
mayswi01

```
select B.playerid
from batting B, halloffame H
where inducted='Y' and H.playerid=B.playerid
group by playerid
order by sum(homeruns) desc
limit 3;
```

b) what 3 players have the highest cumulative total number of homeruns and have NOT been inducted into the hall of fame?

playerid
murphda05
mcgwima01
cepedor01

```
select B.playerid
from batting B, halloffame H
where inducted='N' and H.playerid=B.playerid
group by playerid
order by sum(homeruns) desc
limit 3;
```

8. what player had the highest batting average from 1930 to 1940 (inclusively) with a cumulative total of hits over 1800 and has NOT been elected into the hall of fame? batting average = hits/atbats

average	playerid
0.3515	terrybi01

```
select sum(b.hits)/sum(b.atbats) as average, b.playerid
from batting b join halloffame h on b.playerid=h.playerid
where b.yearid>=1930 and b.yearid<=1940 and h.inducted = 'n'
group by b.playerid
having sum(b.hits)>1800
order by average desc
limit 1;
```

9. what player averaged the highest number of walks from 1900-1920 (inclusively) with over 500 atbats?

playerid
hamilbi01

```
create view lalala as (select sum(b.walks) summation, b.playerid, b.yearid
from batting b
where b.yearid>=1900 and b.yearid<=1920 and b.atbats>500
group by b.playerid, b.yearid);

select playerid
from (select avg(summation) average, playerid
from lalala
group by playerid) as temp
where temp.average in (select max(average)
from (select avg(summation) average, playerid
from lalala
group by playerid) as temp2);
```

10. write a query that uses a non equi join and explain what you are doing.

```
SELECT b.playerID, SUM(b.Hits) AS sum_hits
FROM batting b join (select playerid,min(yearid)as minYear
from halloffame
group by playerid) as f
WHERE b.playerID= f.playerID and b.yearID<f.minYear
GROUP BY b.playerid
ORDER BY sum_hits DESC
```

Get the playerid and their cumulative total number of hits before they are voted into halloffame.

11. write a query that uses an OUTER Join and explain what you are doing.

```
select a.playerid
from players a left join halloffame b on a.playerid = b.playerid
```

Get all players from players table but has no records in the halloffame.