Databases Final Project
Phase I

(1) Who are your team members: Stanley Zheng, Amritpal Singh

(2) Briefly describe your target domain (e.g. a world geographic database) :
Baseball Databases

- Players, Player Statistics, Managers, Teams, Team statistics, team schedules

(3) Give a reasonably comprehensive and representative list of the kinds of English questions you would like your system to be able to answer (minimum 15). For example, “Compute the mean literacy rate for countries with a per capita income of less than $400/year, grouped by continent.” Please note that these queries are not the only thing you will need to support, just some basic objectives to help focus your design choices.

1. Which players have played for a team with a player that played for a team that Yogi Berra played for?
2. Which team has the highest average home runs per game and what is that average?
3. What is the average number of strikeouts per game in 1998, grouped by teams with less than 10 wins that year?
4. List the players with more wins than losses on their record and has played for only 1 team.
5. Of the players who are still alive, how many players were born in the same state as where another player died?
6. List the manager(s) who have managed the most distinct players that bat left handed no matter which team they were on.
7. List the games played in 1963 in Maryland that ended in a tie as well as the teams that played those games.
8. Compute the mean salary between the years 1990-2000 for players whose last name is Smith, grouped by teams.
9. List the players and their birth cities who have played in and won a game in their birth city.
10. List the teams and their managers at the time of the teams who have won the World Series in the same year where they lost more games than they won.
11. How many total strikeouts have the Baltimore Canaries thrown across all their games, grouped by managers?
12. List the players who played on the same team as a player who later became a manager.
13. List the players that have won in a game against a team Yogi Berra was on and also won a game against a team Jackie Robinson was on.

14. List the players who have played in every position at least once.
15. How many games were played in Pennsylvania in which the home team lost?
16. How many teams have won more than 5 World Series?
17. What team won the World Series in 1967?
18. List the teams that Babe Ruth has played on.
19. List all players with more than 10 home runs during the 1954 season.
20. List all players who played with Babe Ruth who had more than 10 home runs during their season with Babe Ruth.

Some Questions we want done by end of week:
Players Questions

- Who has ever played with a certain *player (fname, lname)*
    - SELECT p1.nameFirst, p1.nameLast
      FROM Players p1, Players p2, PlaysFor pf1, PlaysFor pf2
      WHERE p1.retroID = pf1.retroID AND
            p2.retroID = pf2.retroID AND
            pf1.teamID = pf2.teamID AND
            p2.nameFirst = fname AND
            p2.nameLast = lname
- Players with at least a certain amount of *homeruns*, *triples*, *doubles...*
    - SELECT p.nameFirst, p.nameLast
      FROM Players p, Batting b
      WHERE p.playerID = b.playerID AND
            b.HR >= homeruns AND
            b.3B >= triples AND
            b.2B >= doubles
      (optional conditionals depending on which fields are inputted)
- Stats for a certain *player(fname, name)* with multiple bar graphs, one for each year
    - SELECT b.yearID, b.teamID, b.lgID, b.G, b.R, b.H, b.2B, b.3B, b.HR
      (more stats if you want)
      FROM Players p, Batting b
      WHERE p.nameFirst = fname AND
            p.nameLast = lname AND
            p.playerID = b.playerID
- Birth city/state/country for *player (fname, lname)*
    - SELECT p.birthCity, p.birthState, p.birthCountry
      FROM Players p
      WHERE p.nameFirst = fname AND
            p.nameLast = lname

- Players with a certain *hit percentage*
- Player search by *isAlive, batting hand, birthplace* (these fields are optional)
    - SELECT p.nameFirst, p.nameLast
      FROM Players p
      WHERE p.bats =
- List all teams a *player* has played on

Teams Questions

- How many games did a *team* win in a certain *year*
- Team wins by *year* with graph
- Members of team by *year*
- Team above certain *average hit percentage*

(4) Design and show a relational data model that you plan to use for your system, with a preliminary implementation in standard SQL data-definition-language syntax. This specification should include appropriate primary key, foreign key and domain specifications for each relation/attribute, as well as the not null constraint when appropriate. You may also find it useful, but not required, to create a few insert-into statements that populate your schema designs with representative values (both to document your choices and to exercise them. You are welcome to change and augment your design and its specification by Phase II, but any time investment now will reduce effort later.

Tables
Players
Managers
Batting (an entry per player per year)
Pitching (an entry per player per year)
Fielding (an entry per player per year)
PlaysFor
Games

```
DROP TABLE IF EXISTS Players;
CREATE TABLE Players (
        playerID varchar(9) NOT NULL,
        birthYear int(11),
        birthMonth int(11),
        birthDay int(11),
        birthCountry varchar(255),
```

```
        birthState varchar(255),
        birthCity varchar(255),
        deathYear int(11),
        deathMonth int(11),
        deathDay int(11),
        deathCountry varchar(255),
        deathState varchar(255),
        deathCity varchar(255),
        nameFirst varchar(255),
        nameLast varchar(255),
        weight int(11),
        height int(11),
        bats varchar(255),
        throws varchar(255),
        PRIMARY KEY (playerID)
)
INSERT INTO Players VALUES
('aardsda01',1981,12,27,'USA','CO','Denver',NULL,NULL,NULL,NULL,NULL,NULL,'Davi
d','Aardsma",215,75,'R','R')
INSERT INTO Players VALUES
('aaronto01',1939,8,5,'USA','AL','Mobile',1984,8,16,'USA','GA','Atlanta','Tommie','Aaron',
190,75,'R','R')



DROP TABLE IF EXISTS Managers;
CREATE TABLE Managers (
        playerID varchar(9),
        yearID smallint(6) NOT NULL,
        teamID char(3) NOT NULL,
        inseason smallint(6) NOT NULL,
        G smallint(6),
        W smallint(6),
        L smallint(6),
        teamRank smallint(6),
        plyrMgr varchar(1),
        PRIMARY KEY playerID (playerID, yearID),
        CONSTRAINT managers_fk FOREIGN KEY (teamID) REFERENCES teams
       (teamID),
)
```

```
INSERT INTO Managers VALUES ('wrighha01',1871,'BS1',1,31,20,10,3,'Y')
INSERT INTO Managers VALUES ('woodji01',1871,'CH1',1,28,19,9,2,'Y')


DROP TABLE IF EXISTS Teams;
CREATE TABLE Teams (
        yearID smallint(6) NOT NULL,
        teamID char(3) NOT NULL,
        teamRank smallint(6),
        G smallint(6),
        Ghome smallint(6),
        W smallint(6),
        L smallint(6),
        DivWin varchar(1),
        WCWin varchar(1),
        LgWin varchar(1),
        WSWin varchar(1),
        R smallint(6),
        AB smallint(6),
        H smallint(6),
        2B smallint(6),
        3B smallint(6),
        HR smallint(6),
        BB smallint(6),
        SO smallint(6),
        SB smallint(6),
        CS smallint(6),
        HBP smallint(6),
        SF smallint(6),
        RA smallint(6),
        ER smallint(6),
        ERA double,
        CG smallint(6),
        SHO smallint(6),
        SV smallint(6),
        IPouts int(11),
        HA smallint(6),
        HRA smallint(6),
        BBA smallint(6),
        SOA smallint(6),
```

```
        E int(11),
        DP int(11),
        FP double,
        name varchar(50),
        park varchar(255),
        attendance int(11),
        BPF int(11),
        PPF int(11),
        PRIMARY KEY teamID (teamID, yearID),
)
INSERT INTO Teams VALUES
(1871,'BS1',3,31,NULL,20,10,NULL,NULL,'N',NULL,401,1372,426,70,37,3,60,19,73,16,
NULL,NULL,303,109,3.55,22,1,3,828,367,2,42,23,243,24,0.8340000000000001,'Bosto
n Red Stockings','South End Grounds I',NULL,103,98)
INSERT INTO Teams VALUES
(1871,'CH1',2,28,NULL,19,9,NULL,NULL,'N',NULL,302,1196,323,52,21,10,60,22,69,21,
NULL,NULL,241,77,2.76,25,0,1,753,308,6,28,22,229,16,0.8290000000000001,'Chicag
o White Stockings','Union Base-Ball Grounds',NULL,104,102)



DROP TABLE IF EXISTS PlayerBatting;
CREATE TABLE PlayerBatting (
  playerID varchar(9) NOT NULL,
  yearID smallint(6) NOT NULL,
  stint smallint(6) NOT NULL,
  teamID char(3),
  G smallint(6),
  G_batting smallint(6),
  AB smallint(6),
  R smallint(6),
  H smallint(6),
  2B smallint(6),
  3B smallint(6),
  HR smallint(6),
  RBI smallint(6),
  SB smallint(6),
  CS smallint(6),
  BB smallint(6),
```

```
  SO smallint(6),
  IBB smallint(6),
  HBP smallint(6),
  SH smallint(6),
  SF smallint(6),
  GIDP smallint(6),
  PRIMARY KEY playerID (playerID, yearID, stint),
  CONSTRAINT batting_fk2 FOREIGN KEY (teamID) REFERENCES Teams (teamID),
  CONSTRAINT batting_fk2 FOREIGN KEY (playerID) REFERENCES Players
(playerID)
)
INSERT INTO PlayerBatting VALUES
('abercda01',1871,1,'TRO',1,NULL,4,0,0,0,0,0,0,0,0,0,NULL,NULL,NULL,NULL,0)
INSERT INTO PlayerBatting VALUES
('addybo01',1871,1,'RC1',25,NULL,118,30,32,6,0,0,13,8,1,4,0,NULL,NULL,NULL,NULL,
0)
DROP TABLE IF EXISTS PlayerPitching;
CREATE TABLE PlayerPitching (
  playerID varchar(9) NOT NULL,
  yearID smallint(6) NOT NULL,
  stint smallint(6) NOT NULL,
  teamID char(3),
  W smallint(6),
  L smallint(6),
  G smallint(6),
  GS smallint(6),
  CG smallint(6),
  SHO smallint(6),
  SV smallint(6),
  IPouts int(11),
  H smallint(6),
  ER smallint(6),
  HR smallint(6),
  BB smallint(6),
  SO smallint(6),
  BAOpp double,
  ERA double,
  IBB smallint(6),
  WP smallint(6),
  HBP smallint(6),
```

```
  BK smallint(6),
  BFP smallint(6),
  GF smallint(6),
  R smallint(6),
  SH smallint(6),
  SF smallint(6),
  GIDP smallint(6),
  PRIMARY KEY playerID (playerID,yearID,stint),
  CONSTRAINT pitching_fk1 FOREIGN KEY (teamID) REFERENCES Teams(teamID),
  CONSTRAINT pitching_fk2 FOREIGN KEY (playerID) REFERENCES
Players(playerID)
)
INSERT INTO PlayerPitching VALUES
('bechtge01',1871,1,'PH1',1,2,3,3,2,0,0,78,43,23,0,11,1,NULL,7.96,NULL,7,NULL,0,146
,0,42,NULL,NULL,NULL)
```

(5) Submit a set of SQL statements that will implement a representative sample of your target queries, including some of the more interesting or challenging cases. This is primarily to get you to think about your design and how it will be exercised as well as any limitations, so focus on queries that would be useful for doing so, rather than creating trivial or non-insightful queries just to fill space.

- What is the average number of strikeouts in 1998, grouped by teams with less than 10 wins that year?
    - SELECT t.teamID, AVG(t.SO)
      FROM Teams t
      WHERE t.yearID = 1998 AND t.W < 10
      GROUP BY t.teamID
- Which players have played for a team with a player that played for a team that Yogi Berra played for?
    - SELECT p1.nameFirst, p1.nameLast
      FROM Players p1, Players p2, Players yb, PlaysFor pf1, PlaysFor pf2, PlaysFor pf3, PlaysFor pf4
      WHERE pf1.playerID = p1.playerID AND
            pf2.playerID = p2.playerID AND
            pf1.team = pf2.team AND
            pf3.playerID = p2.playerID AND

pf4.playerID = yb.playerID AND
pf3.team = pf4.team

- How many teams have won more than 5 World Series?
    - SELECT WSWin
      FROM Teams
      WHERE WSWin > 5
      GROUP BY teamID
- What team won the World Series in 1967?
    - SELECT name
      FROM Teams
      WHERE yearID = 1967
- List the teams that Babe Ruth has played on.
    - SELECT t.name
      FROM Teams AS t, Players AS p, PlayerBatting AS pb
      WHERE p.playerid = pb.playerid AND t.teamid = pb.teamid AND
      p.nameFirst = Babe AND p.nameLast = Ruth
- List all players with more than 10 home runs during the 1954 season.
    - SELECT p.nameFirst p.nameLast
      FROM Players AS p, PlayerBatting as pb
      WHERE p.playerID = pb.playerID AND HR > 10

(6) Provide a plan for how you will load the database with values. – If you plan to extract/import data from on-line sources, briefly describe what are the sources (e.g. personal data, or provide URL's) and what are any format conversion issues you expect to encounter. – If you plan to input your data primarily through a web or form-based interface, briefly describe this interface and the issues involved.

Lahman Baseball Database: http://www.seanlahman.com/baseball-archive/statistics/
Retrosheet: https://www.retrosheet.org/
Baseball Reference: https://www.baseball-reference.com/

For this baseball database project we plan to extract the data from online sources. Our main source that we are looking at is the Lahman Baseball Database. The Lahman Baseball Database contains complete batting and pitching statistics from 1871 to 2019, plus fielding statistics, standings, team stats, managerial records, post-season data, and more. For our purposes this database should be more than enough to answer all of our questions. Furthermore, the information is the Lahman Database can be downloaded in SQL Lite, MySQL, and Excel so we have options in terms of how we plan on extracting the data. Some format conversion issues we might encounter include having

information in our tables that we might not need. So we would potentially need to prune the data beforehand to get rid of extraneous fields. Another issue might be, if we download the data using Excel, is learning how to convert data from Excel into SQL. I believe this shouldn't be too much of a problem since conversion of data from Excel to SQL seems to be from my experience a common data science task.

(7) Very briefly describe the form/type of output or result you plan to generate or any special user interface issues (e.g. views) that you plan to implement.

We expect to generate a form/type of output where a user can select what they are looking for and insert information according to that and they will then be taken to another page where they can view the information in table form. For example lets say a user wants to see all teammates of Yogi Berra then they can select find teammates and be taken to another page where they can enter the name of the player they want to find all teammates of. From there they will be taken to another page where all of this information is displayed in table form. Some views that we plan to generate are views with a combination of information from Players table and Batting Table, Players Table and Pitching Table, and Players table and Teams table.

(8) What are the specialized/advanced topics you plan to focus on in your database design? Examples include: – security (e.g. banking) – object-oriented or distributed database design/implementation issues – advanced SQL topics (triggers, cursors, JDBC, etc.) – optimization/tuning – data mining – complex data extraction issues from online data sources – natural language interfaces – particularly advanced GUI form interface and/or report generation

- data mining (major)
    - We plan on having a focus on data mining for this project. We plan to mine our data from online sources with forms of data inputs such as Excel tables and MySQL tables that we can download from baseball statistics websites. Furthermore, we can also mine data from websites such as Baseball Reference where we can take the tabular data from these websites and input it into Excel and from there create MySQL tables from this Excel data.
- particularly advanced GUI form interface (minor)
    - For this after talking with Professor Yarowsky we will be implementing some form of graphs for player statistics. Baseball statistics can be very well represented and compared in graph form. So the output of some inputs comparing multiple players can be outputted through graphs.