1. Load batting.csv into a mysql in a database batdb and table batting

##MySQL Prompt

$ mysql -uroot -pcloudera

#create db:

create database batdb;

#switch to database

use batdb

#create table batting:

CREATE TABLE batting (

playerID varchar (20),

yearID int,

stint int,

teamID varchar (20),

lgID varchar (20),

G int,

G\_batting int,

AB int,

R int,

H int,

2B int,

3B int,

HR int,

RBI int,

SB int,

CS int,

BB int,

SO int,

IBB int,

HBP int,

SH int,

SF int,

GIDP int,

G\_old int

);

#load table from csv:

LOAD DATA INFILE '/home/cloudera/Desktop/Batting.csv'

INTO TABLE batting

FIELDS TERMINATED BY ','

ENCLOSED BY '"'

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

2. Sqoop the details into hdfs.

#sqoop into HDFS

sqoop import --connect jdbc:mysql://localhost/batdb --username root --password cloudera --table batting --m 1 --target-dir /batting

3. Sqoop the details into hive.

##Hive prompt

$ hive

#create db in hive

create database sqoop\_db;

#sqoop into hive in command line

$ sqoop import --connect jdbc:mysql://localhost:3306/batdb --username root --password cloudera --split-by yearID --table batting --target-dir /batting\_hive\_3 --fields-terminated-by "," --hive-import --create-hive-table --hive-table sqoop\_db.batting;

5. Implement a Hive script to

a) Find the total count of player details with "david"

use sqoop\_db;

select count(\*) from batting where playerID like "david%";

b) Create a patition on the TEAMID

c) Create 3 buckets on the partition.

create external table batting\_part(playerID string,

yearID int,

stint int,

lgID string,

G int,

G\_batting int,

AB int,

R int,

H int,

B2 int,

B3 int,

HR int,

RBI int,

SB int,

CS int,

BB int,

SO int,

IBB int,

HBP int,

SH int,

SF int,

GIDP int,

G\_old int)

partitioned by (teamID string)

clustered by (lgID) INTO 3 buckets

row format delimited

fields terminated by ','

stored as textfile;

create external table batting\_hive(playerID string,

yearID int,

stint int,

teamID string,

lgID string,

G int,

G\_batting int,

AB int,

R int,

H int,

B2 int,

B3 int,

HR int,

RBI int,

SB int,

CS int,

BB int,

SO int,

IBB int,

HBP int,

SH int,

SF int,

GIDP int,

G\_old int)

row format delimited

fields terminated by ','

stored as textfile;

from batting\_hive bat INSERT OVERWRITE TABLE batting\_part PARTITION(teamID)

select bat.playerID,bat.yearID,

bat.stint ,bat.teamID,bat.lgID ,

bat.G ,bat.G\_batting ,

bat.AB ,bat.R ,bat.H ,

bat.B2 ,bat.B3 ,

bat.HR ,bat.RBI ,

bat.SB ,bat.CS ,bat.BB ,

bat.SO ,bat.IBB, bat.HBP,

bat.SH ,bat.SF,bat.GIDP,

bat.G\_old

DISTRIBUTE BY teamID;

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/Batting.csv' OVERWRITE INTO TABLE batting\_hive;

d) Extract the details on player "aaronha01"

select \* from batting\_part where playerID='aaronha01';

e) Find the count of teams

select count(distinct(teamID)) from batting\_hive;

4. Implement a PIG script to

##load data to the pig prompt -->> pig

batting\_list = LOAD '/home/cloudera/Batting.csv' USING PigStorage(',') as (playerID:chararray,

yearID:int,

stint:int,

teamID:chararray,

lgID:chararray,

G:int,

G\_batting:int,

AB:int,

R:int,

H:int,

B2:int,

B3:int,

HR:int,

RBI:int,

SB:int,

CS:int,

BB:int,

SO:int,

IBB:int,

HBP:int,

SH:int,

SF:int,

GIDP:int,

G\_old:int);

----

a) Find the total count of participation of G 112

----

count\_g = FILTER batting\_list BY G == 112;

group\_count\_g = GROUP count\_g All;

total\_count = foreach group\_count\_g Generate COUNT(count\_g.G);

dump total\_count;

### --->> 209

----

b) Find the player details with "david"

----

david = Filter batting\_list by(playerID MATCHES 'david.\*');

dump david;

Store david into 'home/cloudera/Desktop/assign\_pigresults';

----

c) Find the average count of "NL"

----

NL\_filter = Filter batting\_list by lgID =='NL';

NL\_Group = Group NL\_filter All;

NL\_avg = foreach NL\_Group Generate AVG(NL\_filter.G\_batting);

DUMP NL\_avg;

### --->> 51.957273064709035

----

d) Find the count of teams

----

team\_count = GROUP batting\_list by teamID;

team\_group = GROUP team\_count All;

result\_count = Foreach team\_group Generate COUNT(team\_count);

dump result\_count;

### --->> 149

9. Using hive,partition by year. Then, find the year wise count of participants,

find the total votes got by the players.

create table halloffame(hofID STRING,

yearid INT,

votedBy STRING,

ballots INT,

needed INT,

votes INT,

inducted STRING,

category STRING,

needed\_note STRING)

row format delimited fields terminated by ','

stored as textfile;

LOAD DATA LOCAL INPATH '/home/cloudera/Desktop/HallOfFame.csv' into table halloffame;

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.exec.dynamic.partition=true;

set hive.enforce.bucketing=true;

create table halloffame\_part(hofID STRING,

votedBy STRING,

ballots INT,

needed INT,

votes INT,

inducted STRING,

category STRING,

needed\_note STRING)

partitioned by(yearid INT)

row format delimited fields terminated by ','

lines terminated by '\n';

from halloffame hof INSERT OVERWRITE TABLE halloffame\_part PARTITION(yearid) select hof.hofID,

hof.votedBy,

hof.ballots,

hof.needed,

hof.votes,

hof.inducted,

hof.category,

hof.needed\_note,

hof.yearid DISTRIBUTE BY yearid;

select yearid, count(hofid) from halloffame\_part group by yearid;

select hofid, sum(votes) from halloffame\_part group by hofid;