CLASS PROJECTS – BIG DATA 2019

CCBD PROJECT TIMELINES

- Mid Review
  - Talk to your guides by Oct 31st
  - Prepare a presentation on goals in discussion with Guide. Agree on goals to be achieved and submit the ppt.
  - Presentation format is shared on Piazza.
- Final Project Submission
  - Nov 25-Nov 29
  - Schedule will be put up, you can select your slot.

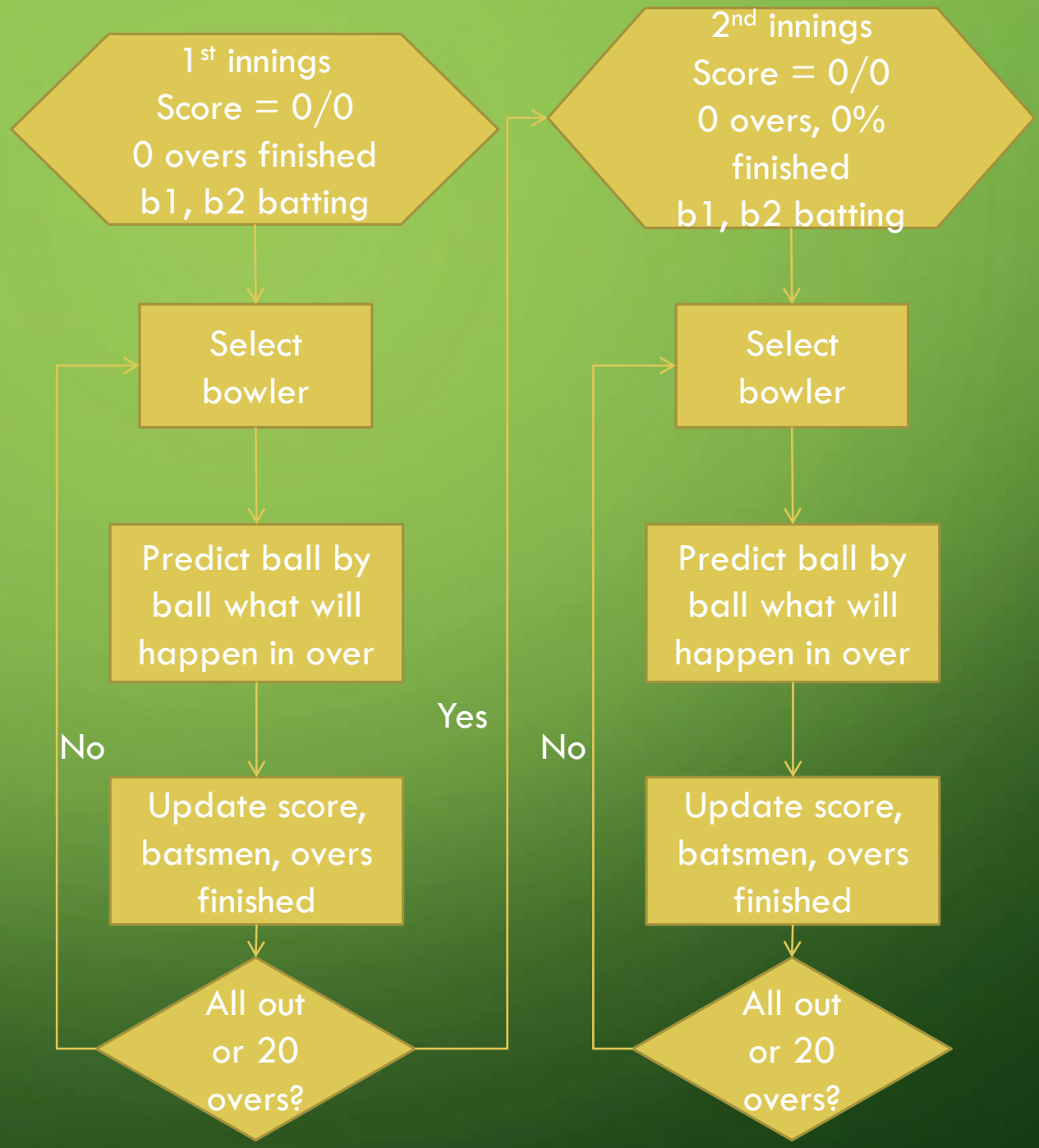
AVAILABLE PROJECTS

- Analysis based
  - IPL Analysis
- Coding based
  - SQL queries using Map-Reduce
- Note: please write original code. We will run your code through the plagiarism checker

IPL ANALYSIS – CLASS PROJECT

OVERVIEW

- Given
  - 2 cricket teams
  - Batting order
  - Bowling order
  - Who bats first
- Predict the score
- How?
  - Flowchart at right



PREDICTING OUTCOME OF OVER

- Predict ball by ball
  - Ignore extras for now
  - Handle wickets separately
  - Ignore stage of the game
    - 1st innings – how many overs are complete
    - 2nd innings – overs complete, % target complete
- We need to calculate p(0, 1, 2, 3, 4, 6 runs scored), for each ball
- This depends upon
  - Who is batting
  - Who is bowling

ESTIMATING PROBABILITIES

- Maximum Likelihood Estimator
  - P(event) = (no. of times event occurred)/(total number of events)
- Example
  - Suppose McCleneghan bowls 40 balls to Shikhar Dhawan
  - 4 times Shikhar Dhawan hits a 4
  - P(hitting 4) = 0.1

SIMPLE APPROACH (MAY NOT WORK)

1. Go over all the games
2. If we find a game where
    1. Shikhar Dhawan is batting
    2. McCleneghan bowling
    3. 4 runs are scored
3. Add 1 to the number of times 4 is hit
4. Calculate probabilities

PROBLEM WITH SIMPLE APPROACH

- Many times, in IPL, we may need combinations we haven't seen before or have seen rarely
  - If it's a rare event, probability calculation may not be accurate

SOLUTION 1

- Cluster (group together) similar batsmen into say 10 groups
- Also cluster similar bowlers
- Calculate probability of n runs when
  - Batsman of group A is batting
  - Bowler of group X is bowling

CLUSTERING DETAILS

- Download from cricinfo.com
  - Player vs player statistics for all T20s in 2016
  - Player profiles
- Use these player profiles to cluster batsmen, bowlers
- From the player vs player statistics, calculate the probability of runs being scored in over
- Clustering code for k-means should be written by student using Map Reduce.
  - No marks for using standard code.

SOLUTION 2

- Use collaborative filtering to identify the probabilities for the batsman-bowler combinations not seen before.
- Use mllib (https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html)

APPROACH TO COMPUTING #RUNS SCORED PER BALL

- Determine batsman-bowler combination (or batsman cluster-bowler cluster combination)
  - Use cluster if enough data is not there about specific batsman bowler
  - Use both solution1 and 2 to compute probabilities
- Generate a random number
- Look up cumulative probability distribution of runs scored to see where the random number falls
  - Use that as the total runs scored.

- Example
  - If random number val on first ball = 0.91 then look up table below.
  - 0.85-1.0 Is a 6, so a 6 is scored

Runs	P(Runs)	CumProb(Runs)
0	0.1	0.1
1	0.2	0.3
2	0.2	0.5
3	0.1	0.6
4	0.25	0.85
6	0.15	1.0

PLAYER VS PLAYER INFO

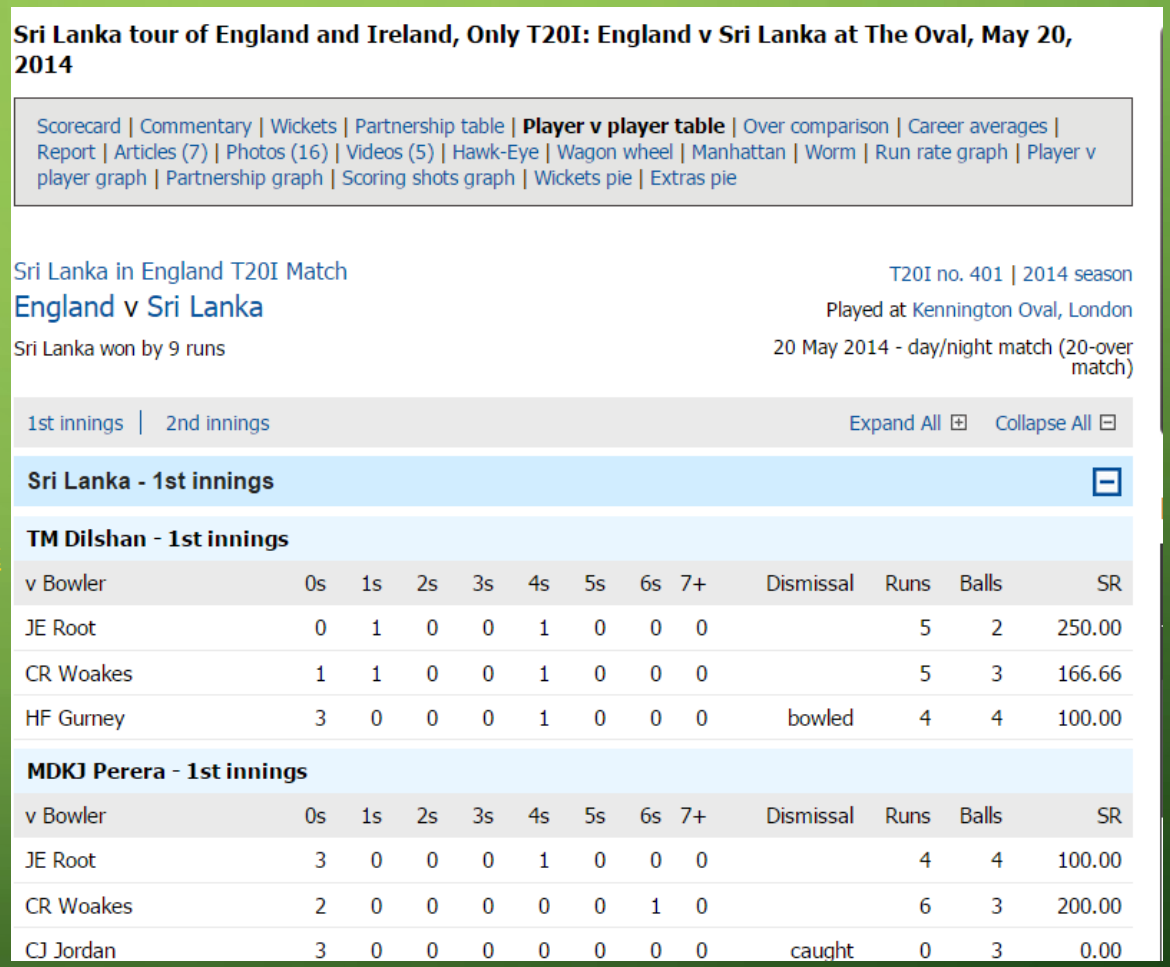
Get data from cricsheets

Compute player vs player profiles by parsing the sheets.

Or

Use the tabulated player statistics from ESPN Cricinfo to obtain important data points such as Strike Rate, Average etc. An example may be found at - http://stats.espncricinfo.com/indian-premier-league-2016/engine/records/averages/batting.html?id=11001;type=tournament

Maybe write a MR program or a python program – left to you.



Sri Lanka tour of England and Ireland, Only T20I: England v Sri Lanka at The Oval, May 20, 2014

Scorecard | Commentary | Wickets | Partnership table | **Player v player table** | Over comparison | Career averages | Report | Articles (7) | Photos (16) | Videos (5) | Hawk-Eye | Wagon wheel | Manhattan | Worm | Run rate graph | Player v player graph | Partnership graph | Scoring shots graph | Wickets pie | Extras pie

Sri Lanka in England T20I Match
England v Sri Lanka
Sri Lanka won by 9 runs

T20I no. 401 | 2014 season
Played at Kennington Oval, London
20 May 2014 - day/night match (20-over match)

1st innings | 2nd innings — Expand All ⊞ Collapse All ⊟

Sri Lanka - 1st innings

TM Dilshan - 1st innings

v Bowler	0s	1s	2s	3s	4s	5s	6s	7+	Dismissal	Runs	Balls	SR
JE Root	0	1	0	0	1	0	0	0		5	2	250.00
CR Woakes	1	1	0	0	1	0	0	0		5	3	166.66
HF Gurney	3	0	0	0	1	0	0	0	bowled	4	4	100.00

MDKJ Perera - 1st innings

v Bowler	0s	1s	2s	3s	4s	5s	6s	7+	Dismissal	Runs	Balls	SR
JE Root	3	0	0	0	1	0	0	0		4	4	100.00
CR Woakes	2	0	0	0	0	0	1	0		6	3	200.00
CJ Jordan	3	0	0	0	0	0	0	0	caught	0	3	0.00

PLAYER PROFILE

http://www.espncricinfo.com/ci/content/player/50710.html

Full name Kumar Chokshanada Sangakkara

Born October 27, 1977, Matale

Current age 37 years 302 days

Major teams Sri Lanka, Asia XI, Central Province, Colombo District Cricket Association, Deccan Chargers, Durham, ICC World XI, Jamaica Tallawahs, Kandurata, Kandurata Maroons, Kings XI Punjab, Marylebone Cricket Club, Nondescripts Cricket Club, Sunrisers Hyderabad, Warwickshire

Playing role Wicketkeeper batsman

Batting style Left-hand bat

Bowling style Right-arm offbreak

Fielding position Wicketkeeper

Like 17k

Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	134	233	17	12400	319	57.40	22882	54.19	38	52	1491	51	182	20
ODIs	404	380	41	14234	169	41.98	18048	78.86	25	93	1385	88	402	99
T20Is	56	53	9	1382	78	31.40	1156	119.55	0	8	139	20	25	20
First-class	235	387	28	18134	319	50.51			53	77			352	33
List A	509	481	51	18389	169	42.76			35	115			508	124
Twenty20	164	158	17	4214	94	29.88	3355	125.60	0	24	457	76	99	41

Bowling averages

	Mat	Inns	Balls	Runs	Wkts	BBI	BBM	Ave	Econ	SR	4w	5w	10
Tests	134	4	84	49	0	-	-	-	3.50	-	0	0	0
ODIs	404	-	-	-	-	-	-	-	-	-	-	-	-
T20Is	56	-	-	-	-	-	-	-	-	-	-	-	-
First-class	235		246	150	1	1/13		150.00	3.65	246.0		0	0
List A	509	-	-	-	-	-	-	-	-	-	-	-	-
Twenty20	164	-	-	-	-	-	-	-	-	-	-	-	-

Career statistics

WICKETS

- From statistics, calculate probability of batsman being out
- Example
  - Suppose A&B are batting
  - X,Y,Z are bowling
  - P(A out when X bowling) = 0.04
  - P(B out when Y bowling) = 0.06
  - P(A out when Z bowling) = 0.08

- Fall of wickets
  - Ball 1: X bowls to A; p (A is not out) = 0.96
  - …
  - Ball 6: X bowls; p(A is not out) = 0.78
  - Ball 7: Y bowls to B; p(B is not out) = 0.94
  - Ball 12: Y bowls; p(B is not out) = 0.69
  - Ball 13: Z bowls to A; p(A is not out) = 0.78 x 0.92 = 0.72
  - Ball 17: Z bowls; p(A is not out) = 0.51
  - Ball 18: Z bowls; p(A is not out) = 0.47 < 0.5
    - Assume wicket falls on 18th ball

STEPS

- Step 1: Load player data into HDFS, cluster batsmen, bowlers into groups (5 marks)
- Step 2: Load group vs group statistics, simulate match using the clustering approach(10 marks)
- Step 3: Simulate match score using the recommendation algorithm approach and compare with Approach using clustering(10 marks)

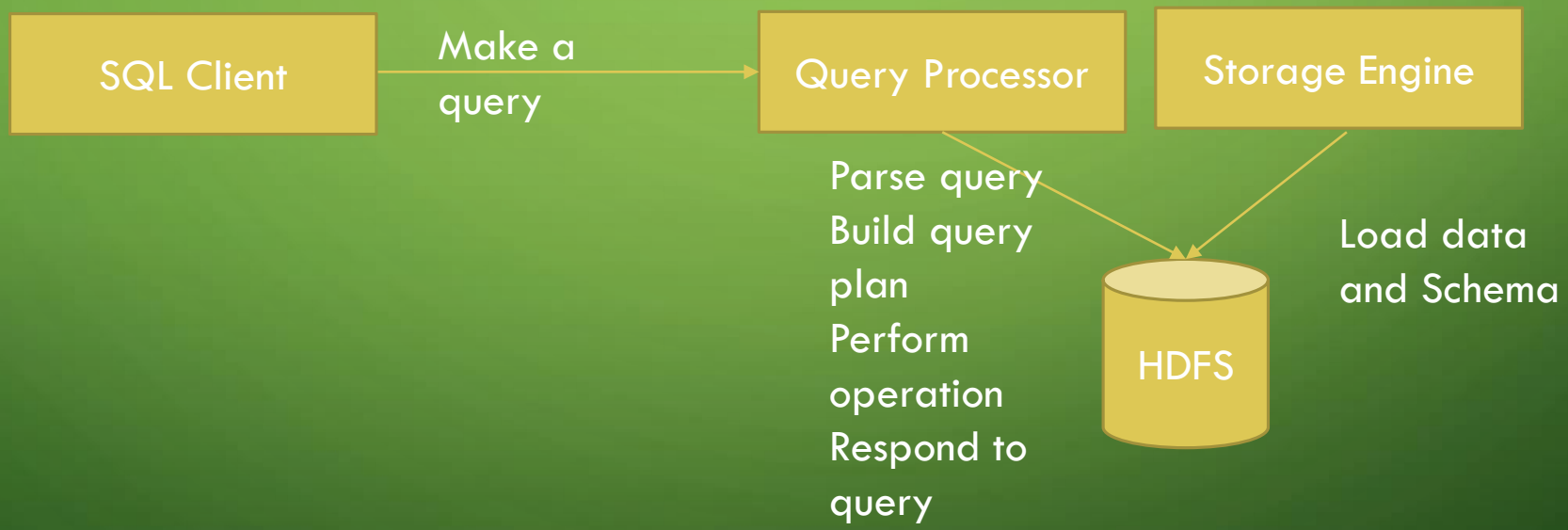
BUILDING A MAPREDUCE BASED SQL ENGINE

BIG DATA 2019– CLASS PROJECT

OBJECTIVE

- Get a practical insight into building a map reduce based sql engines
- Understand the design of a big data storage system

ARCHITECTURE



OPERATIONS SUPPORTED

- Load database
- Delete database
- Select/project
- Aggregate: pick 3 of COUNT, MIN, MAX, SUM, AVG

STEP1: LOADING THE DATABASE

- Load data into database
  - Data is stored in CSV file on HDFS
  - On loading data, also specify the schema
  - Example
    - LOAD database_name/table_name.csv AS (column_name: datatype, column_name: datatype);
    - LOAD bigdata/project_list.csv AS (student_name: string, year: integer, cgpa:integer)
- Design Challenge: where will you store the schema?
- 5 marks

STEP 2: SELECT AND PROJECT QUERIES

- You should be able to parse and run a select/project query
- Take care of standard errors like the column name not found or table not found.
- Example:
  - SELECT column_name FROM database_name/table_name.csv WHERE column_name = value;
- Design Challenge
  - Convert the parsed query to a set of map-reduce jobs – how to define these jobs
  - What happens when the map-reduce jobs do not complete on time?
- 10 marks

STEP 3: AGGREGATE QUERIES

- You should be able to parse and run aggregate queries for any 3 of MIN, MAX, COUNT, AVG, SUM
- Remember that aggregate can happen along with select and project
- Take care of standard errors like summing on a string
- Example:
  - SELECT column_name FROM database_name/table_name.csv WHERE column_name = value;
- Design Challenge
  - Order of operations?
- 10 marks

APPROXIMATE DATES

- Step 1 – Nov 11th
- Step 2 – Nov 18th
- Step 3 – week of Nov 25th.

- These dates are only for guidance. You can complete before that date also. Please talk to one of the teachers to setup your evaluation.
- Plan your work. Other courses will also have projects.