CS313:Big Data

SQL Engine Based on Map-Reduce

Mini-Hive for basic operations

Big Data | 2/12/2019

SNo	Name	USN	Class/Section
1	Mainaki Saraf	PES1201701002	D
2	Kanishk Singh	PES1201700223	D
3	Abhinandan Singla	PES1201701128	B
4	Akash Mukhopadhyay	PES1201700133	G

Introduction

This project is a mini implementation of Hive that performs operations like LOAD dataset, SELECT queries and AGGREGATE_BY (max , min, sum, avg) and DELETE operations on a given dataset. These queries run map-reduce jobs in the background to give the output by performing the desired actions on the provided dataset.

Related work

As this project is a mini-implementation of Hive, which is an opensource software, the original repository of Hive on GitHub was referred. Hive convert all SQL queries to map-reduce jobs and gives the output.

ALGORITHM/DESIGN

The project can implement LOAD, SELECT, AGGREGATE_BY and DELETE operations on a given dataset.

LOAD:
The dataset with its local path is given as input along with the required schema. On this command, in the predefined folder /files in hdfs, a folder with the name of the dataset is created and the csv file is stored in this folder. The schema of the dataset is stored locally and on hdfs in a folder /schema as a text file with the name of the dataset file.

DELETE:
The dataset is deleted from hdfs and the folder with the dataset name in /files is also deleted. The schema text file in /schema is also deleted form hdfs and locally.

SELECT:
This query enables the user to view various columns of the dataset. Single column can be used, multiple columns or all columns (*) can be projected. Conditions can be added to the projection using the WHERE keyword followed by the condition on a given column. Conditions like logical operations (=, >=, <=, !=) for columns can be performed.

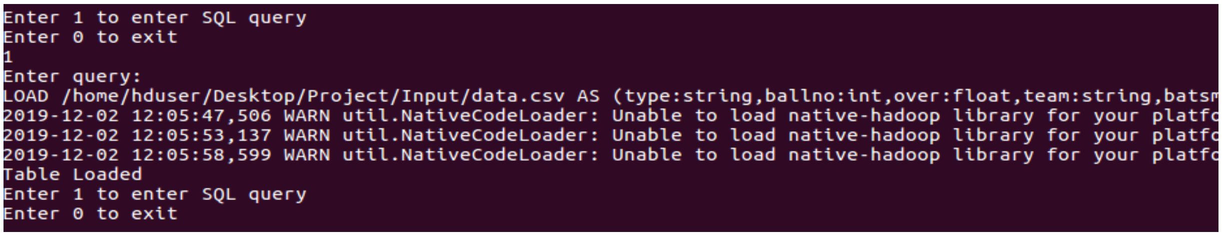
The mapper is responsible for filtering the data based on the column names and the condition required and sends the result to the reducer. The reducer is an identity reducer and prints the values sent by the mapper.

AGGRGEGATE_BY (min, max, sum, avg):
This query is paired up with the SELECT query. It involves the functions: min, max, sum, avg to be performed on the projected column. The column has to be a numeric column for these functions else the query fails. The output returns a statement with the projected column name and the aggregated value based on the function specified.

The mapper selects the rows based on the condition given in the where clause if any and passes the projected columns name and value to the reducer as a key value pair. The reducer performs the desired aggregation on these values and prints the desired output for the aggregation specified for the given column.

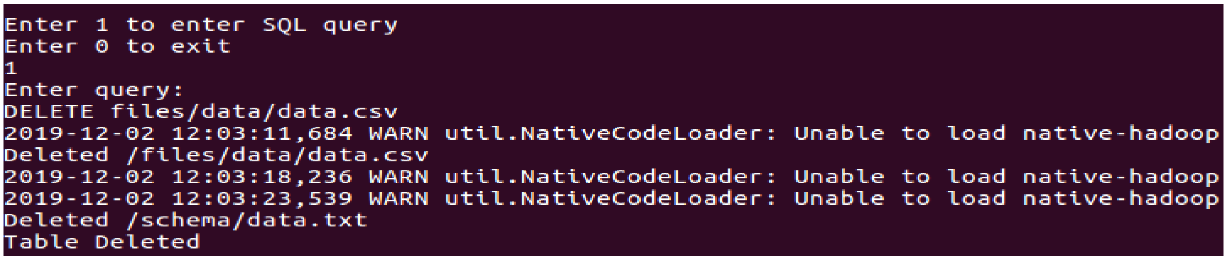
EXPERIMENTAL RESULTS

LOAD:



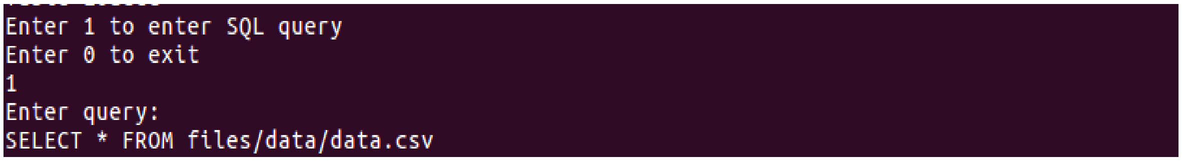
```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
LOAD /home/hduser/Desktop/Project/Input/data.csv AS (type:string,ballno:int,over:float,team:string,batsm
2019-12-02 12:05:47,506 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
2019-12-02 12:05:53,137 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
2019-12-02 12:05:58,599 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platfo
Table Loaded
Enter 1 to enter SQL query
Enter 0 to exit
```

DELETE:

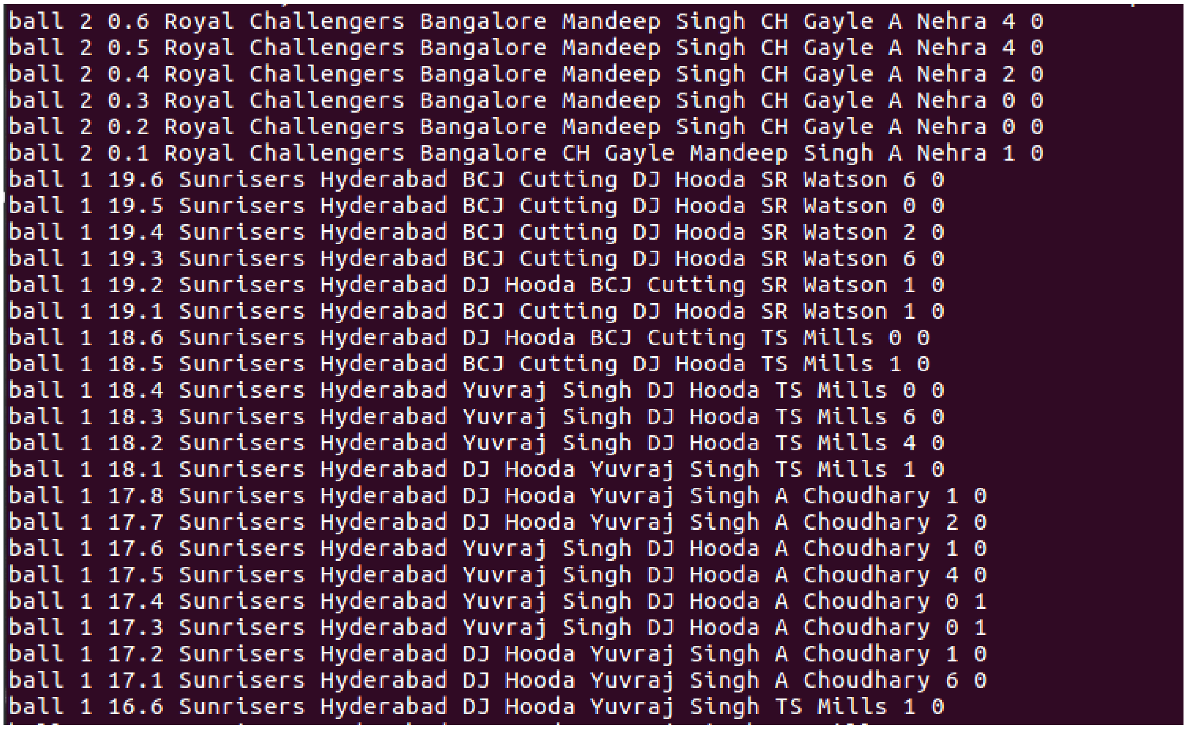


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
DELETE files/data/data.csv
2019-12-02 12:03:11,684 WARN util.NativeCodeLoader: Unable to load native-hadoop
Deleted /files/data/data.csv
2019-12-02 12:03:18,236 WARN util.NativeCodeLoader: Unable to load native-hadoop
2019-12-02 12:03:23,539 WARN util.NativeCodeLoader: Unable to load native-hadoop
Deleted /schema/data.txt
Table Deleted
```

SELECT without WHERE condition:

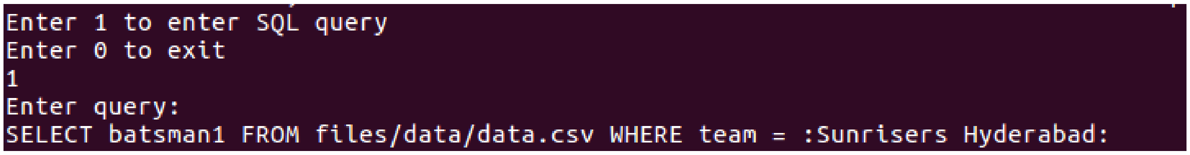


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT * FROM files/data/data.csv
```

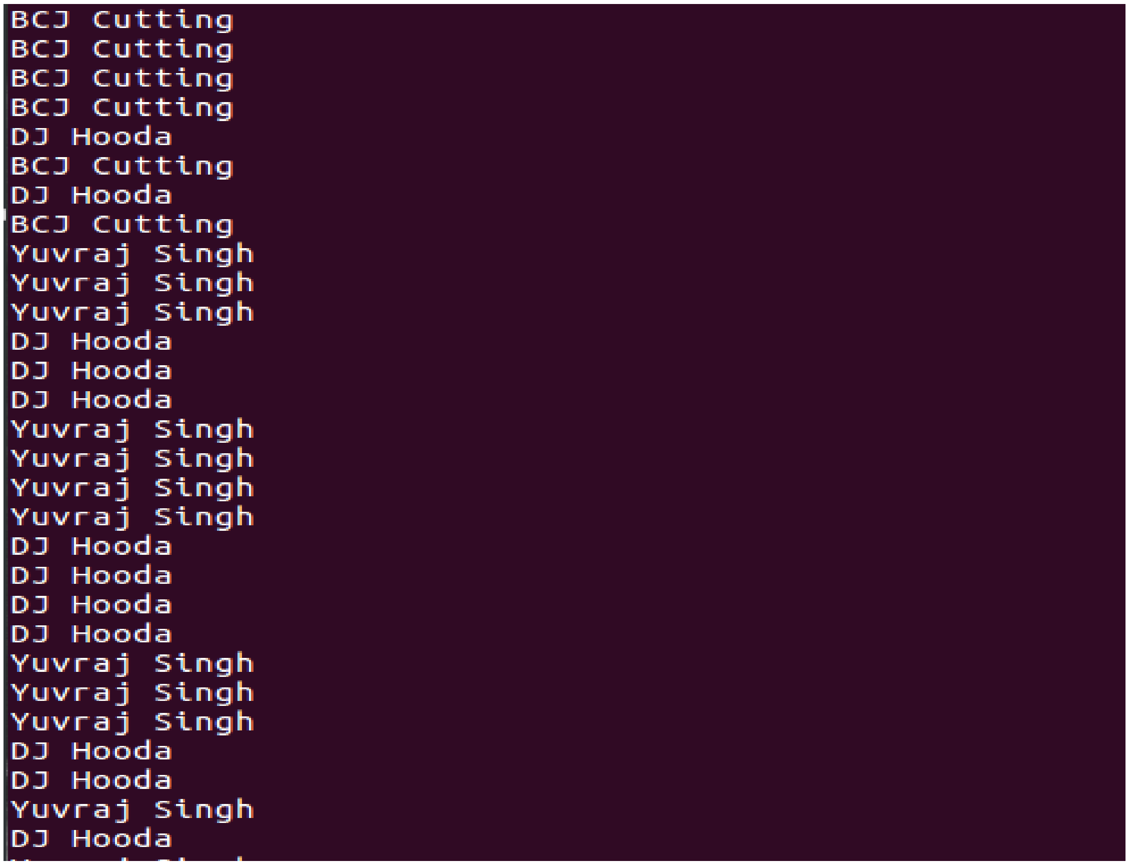


```
ball 2 0.6 Royal Challengers Bangalore Mandeep Singh CH Gayle A Nehra 4 0
ball 2 0.5 Royal Challengers Bangalore Mandeep Singh CH Gayle A Nehra 4 0
ball 2 0.4 Royal Challengers Bangalore Mandeep Singh CH Gayle A Nehra 2 0
ball 2 0.3 Royal Challengers Bangalore Mandeep Singh CH Gayle A Nehra 0 0
ball 2 0.2 Royal Challengers Bangalore Mandeep Singh CH Gayle A Nehra 0 0
ball 2 0.1 Royal Challengers Bangalore CH Gayle Mandeep Singh A Nehra 1 0
ball 1 19.6 Sunrisers Hyderabad BCJ Cutting DJ Hooda SR Watson 6 0
ball 1 19.5 Sunrisers Hyderabad BCJ Cutting DJ Hooda SR Watson 0 0
ball 1 19.4 Sunrisers Hyderabad BCJ Cutting DJ Hooda SR Watson 2 0
ball 1 19.3 Sunrisers Hyderabad BCJ Cutting DJ Hooda SR Watson 6 0
ball 1 19.2 Sunrisers Hyderabad DJ Hooda BCJ Cutting SR Watson 1 0
ball 1 19.1 Sunrisers Hyderabad BCJ Cutting DJ Hooda SR Watson 1 0
ball 1 18.6 Sunrisers Hyderabad DJ Hooda BCJ Cutting TS Mills 0 0
ball 1 18.5 Sunrisers Hyderabad BCJ Cutting DJ Hooda TS Mills 1 0
ball 1 18.4 Sunrisers Hyderabad Yuvraj Singh DJ Hooda TS Mills 0 0
ball 1 18.3 Sunrisers Hyderabad Yuvraj Singh DJ Hooda TS Mills 6 0
ball 1 18.2 Sunrisers Hyderabad Yuvraj Singh DJ Hooda TS Mills 4 0
ball 1 18.1 Sunrisers Hyderabad DJ Hooda Yuvraj Singh TS Mills 1 0
ball 1 17.8 Sunrisers Hyderabad DJ Hooda Yuvraj Singh A Choudhary 1 0
ball 1 17.7 Sunrisers Hyderabad DJ Hooda Yuvraj Singh A Choudhary 2 0
ball 1 17.6 Sunrisers Hyderabad Yuvraj Singh DJ Hooda A Choudhary 1 0
ball 1 17.5 Sunrisers Hyderabad Yuvraj Singh DJ Hooda A Choudhary 4 0
ball 1 17.4 Sunrisers Hyderabad Yuvraj Singh DJ Hooda A Choudhary 0 1
ball 1 17.3 Sunrisers Hyderabad Yuvraj Singh DJ Hooda A Choudhary 0 1
ball 1 17.2 Sunrisers Hyderabad DJ Hooda Yuvraj Singh A Choudhary 1 0
ball 1 17.1 Sunrisers Hyderabad DJ Hooda Yuvraj Singh A Choudhary 6 0
ball 1 16.6 Sunrisers Hyderabad DJ Hooda Yuvraj Singh TS Mills 1 0
```

SELECT with WHERE clause:

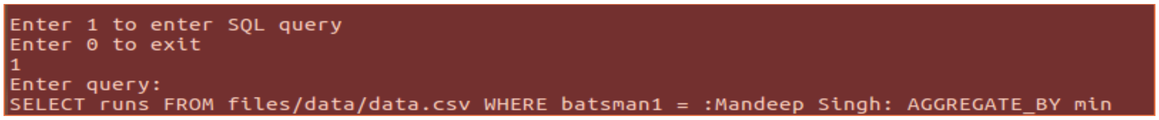


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT batsman1 FROM files/data/data.csv WHERE team = :Sunrisers Hyderabad:
```

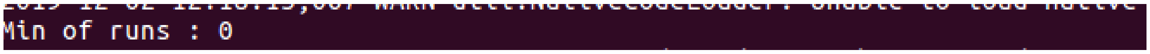


```
BCJ Cutting
BCJ Cutting
BCJ Cutting
BCJ Cutting
DJ Hooda
BCJ Cutting
DJ Hooda
BCJ Cutting
Yuvraj Singh
Yuvraj Singh
Yuvraj Singh
DJ Hooda
DJ Hooda
DJ Hooda
Yuvraj Singh
Yuvraj Singh
Yuvraj Singh
Yuvraj Singh
DJ Hooda
DJ Hooda
DJ Hooda
DJ Hooda
Yuvraj Singh
Yuvraj Singh
Yuvraj Singh
DJ Hooda
DJ Hooda
Yuvraj Singh
DJ Hooda
```

AGGREGATE_BY min:

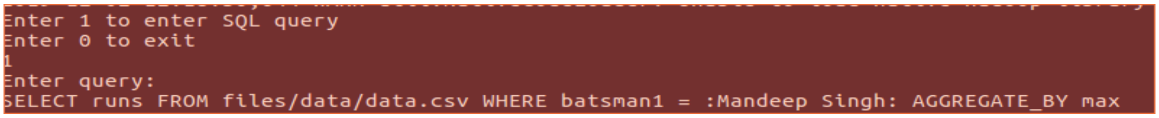


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT runs FROM files/data/data.csv WHERE batsman1 = :Mandeep Singh: AGGREGATE_BY min
```

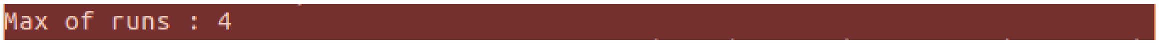


```
Min of runs : 0
```

AGGREGATE_BY max:

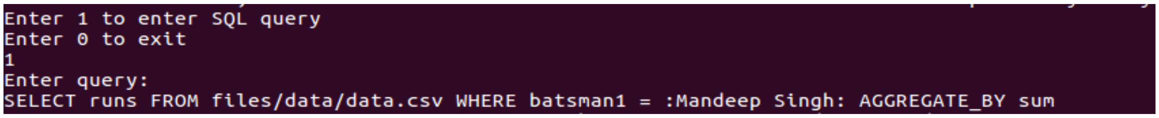


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT runs FROM files/data/data.csv WHERE batsman1 = :Mandeep Singh: AGGREGATE_BY max
```

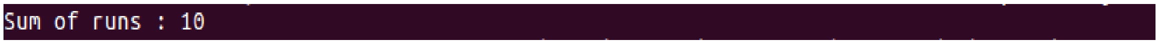


```
Max of runs : 4
```

AGGREGATE_BY sum:

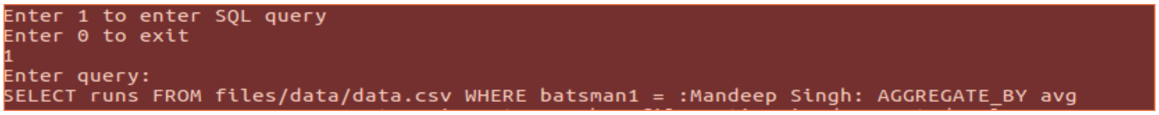


```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT runs FROM files/data/data.csv WHERE batsman1 = :Mandeep Singh: AGGREGATE_BY sum
```

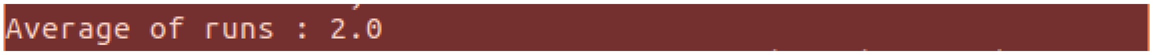


```
Sum of runs : 10
```

AGGREGATE_BY avg:



```
Enter 1 to enter SQL query
Enter 0 to exit
1
Enter query:
SELECT runs FROM files/data/data.csv WHERE batsman1 = :Mandeep Singh: AGGREGATE_BY avg
```



```
Average of runs : 2.0
```

FUTURE ENHANCEMENTS

The project can be further extended to incorporated more complex queries with WHERE condition spanning over multiple columns. More aggregate functions along with the GROUP_BY clause can be included.

REFERENCES

https://github.com/apache/hive
https://pdfs.semanticscholar.org/b361/1f5d72768dbfe2c78030a9433276feb63c04.pdf

EVALUATIONS (Leave this for the faculty)

Date	Evaluator	Comments	Score

CHECKLIST

SNo	Item	Status
1.	Source code documented	
2	Source code uploaded to portal – submission id (only for class project)	
3.	Instructions for building and running the code. Your code must be usable out of the box. Link to your gitlab account (for CCBD project only)	