DDL (Data definition language) for Risk Factor Project

1) Create table ‘geolocation’:

a)

```
CREATE TABLE geolocation
(
truckid string,
driverid string,
event string,
latitude DOUBLE,
longitude DOUBLE,
city string,
state string,
velocity BIGINT,
event_ind  BIGINT,
idling_ind BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

b) Load the ‘geolocation’ tab data into geolocation table.

2) Create table ‘trucks’:

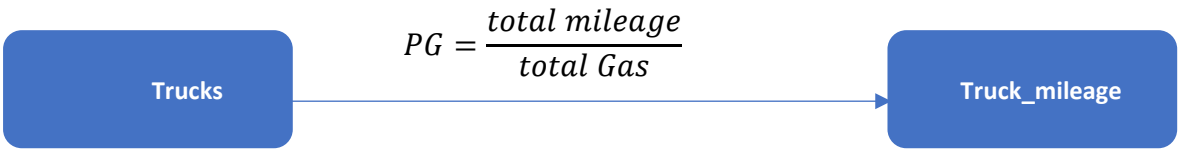
a)

```
CREATE TABLE trucks(driverid string, truckid string, model string, jun13_miles bigint, jun13_gas bigint,
may13_miles bigint, may13_gas bigint, apr13_miles bigint, apr13_gas bigint, mar13_miles bigint,
mar13_gas bigint, feb13_miles bigint, feb13_gas bigint, jan13_miles bigint, jan13_gas bigint, dec12_miles
bigint, dec12_gas bigint, nov12_miles bigint, nov12_gas bigint, oct12_miles bigint, oct12_gas bigint,
sep12_miles bigint, sep12_gas bigint, aug12_miles bigint, aug12_gas bigint, jul12_miles bigint, jul12_gas
bigint, jun12_miles bigint, jun12_gas bigint,may12_miles bigint, may12_gas bigint, apr12_miles bigint,
apr12_gas bigint, mar12_miles bigint, mar12_gas bigint, feb12_miles bigint, feb12_gas bigint, jan12_miles
bigint, jan12_gas bigint, dec11_miles bigint, dec11_gas bigint, nov11_miles bigint, nov11_gas bigint,
oct11_miles bigint, oct11_gas bigint, sep11_miles bigint, sep11_gas bigint, aug11_miles bigint, aug11_gas
bigint, jul11_miles bigint, jul11_gas bigint, jun11_miles bigint, jun11_gas bigint, may11_miles bigint,
may11_gas bigint, apr11_miles bigint, apr11_gas bigint, mar11_miles bigint, mar11_gas bigint,
feb11_miles bigint, feb11_gas bigint, jan11_miles bigint, jan11_gas bigint, dec10_miles bigint, dec10_gas
bigint, nov10_miles bigint, nov10_gas bigint, oct10_miles bigint, oct10_gas bigint, sep10_miles bigint,
sep10_gas bigint, aug10_miles bigint, aug10_gas bigint, jul10_miles bigint, jul10_gas bigint, jun10_miles
bigint, jun10_gas bigint, may10_miles bigint, may10_gas bigint, apr10_miles bigint, apr10_gas bigint,
mar10_miles bigint, mar10_gas bigint, feb10_miles bigint, feb10_gas bigint, jan10_miles bigint, jan10_gas
bigint, dec09_miles bigint, dec09_gas bigint, nov09_miles bigint, nov09_gas bigint, oct09_miles bigint,
oct09_gas bigint, sep09_miles bigint, sep09_gas bigint, aug09_miles bigint, aug09_gas bigint, jul09_miles
bigint, jul09_gas bigint, jun09_miles bigint, jun09_gas bigint, may09_miles bigint, may09_gas bigint,
apr09_miles bigint, apr09_gas bigint, mar09_miles bigint, mar09_gas bigint, feb09_miles bigint,
feb09_gas bigint, jan09_miles bigint, jan09_gas bigint)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

c) Load the data from ‘trucks’ input file into the table “trucks”

3) Create table ‘truck_milaeage’ from existing ‘trucks’ table:

a)



CREATE TABLE truck_mileage AS SELECT truckid, driverid, rdate, miles, gas, miles / gas mpg FROM trucks LATERAL VIEW stack(54, 'jun13',jun13_miles,jun13_gas,'may13',may13_miles,may13_gas,'apr13',apr13_miles,apr13_gas,'mar13',mar13_miles,mar13_gas,'feb13',feb13_miles,feb13_gas,'jan13',jan13_miles,jan13_gas,'dec12',dec12_miles,dec12_gas,'nov12',nov12_miles,nov12_gas,'oct12',oct12_miles,oct12_gas,'sep12',sep12_miles,sep12_gas,'aug12',aug12_miles,aug12_gas,'jul12',jul12_miles,jul12_gas,'jun12',jun12_miles,jun12_gas,'may12',may12_miles,may12_gas,'apr12',apr12_miles,apr12_gas,'mar12',mar12_miles,mar12_gas,'feb12',feb12_miles,feb12_gas,'jan12',jan12_miles,jan12_gas,'dec11',dec11_miles,dec11_gas,'nov11',nov11_miles,nov11_gas,'oct11',oct11_miles,oct11_gas,'sep11',sep11_miles,sep11_gas,'aug11',aug11_miles,aug11_gas,'jul11',jul11_miles,jul11_gas,'jun11',jun11_miles,jun11_gas,'may11',may11_miles,may11_gas,'apr11',apr11_miles,apr11_gas,'mar11',mar11_miles,mar11_gas,'feb11',feb11_miles,feb11_gas,'jan11',jan11_miles,jan11_gas,'dec10',dec10_miles,dec10_gas,'nov10',nov10_miles,nov10_gas,'oct10',oct10_miles,oct10_gas,'sep10',sep10_miles,sep10_gas,'aug10',aug10_miles,aug10_gas,'jul10',jul10_miles,jul10_gas,'jun10',jun10_miles,jun10_gas,'may10',may10_miles,may10_gas,'apr10',apr10_miles,apr10_gas,'mar10',mar10_miles,mar10_gas,'feb10',feb10_miles,feb10_gas,'jan10',jan10_miles,jan10_gas,'dec09',dec09_miles,dec09_gas,'nov09',nov09_miles,nov09_gas,'oct09',oct09_miles,oct09_gas,'sep09',sep09_miles,sep09_gas,'aug09',aug09_miles,aug09_gas,'jul09',jul09_miles,jul09_gas,'jun09',jun09_miles,jun09_gas,'may09',may09_miles,may09_gas,'apr09',apr09_miles,apr09_gas,'mar09',mar09_miles,mar09_gas,'feb09',feb09_miles,feb09_gas,'jan09',jan09_miles,jan09_gas) dummyalias AS rdate, miles, gas;

b) Validate the new tables in the database.
Verify that both the geolocation, trucks and truck_mileage tables are in the default database.

4) Create table avg_mileage from existing trucks_mileage table:

a)

```
CREATE TABLE avg_mileage
AS
SELECT truckid, avg(mpg) avgmpg
FROM truck_mileage
GROUP BY truckid;
```

5) Create table 'drivermileage' from existing 'truck_mileage' table

```
CREATE TABLE DriverMileage
AS
SELECT driverid, sum(miles) totmiles
FROM truck_mileage
GROUP BY driverid;
```

6) Create table 'trucks_mg"

a)

```
CREATE TABLE trucks_mg(driverid string, truckid string, model string, Tdate
string, miles bigint, gas bigint )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");
```

b)

Load 'trucks_mg' input file into 'trucks_mg' table.

7) Create table riskfactor:

a)

```
CREATE TABLE riskfactor_idling
(driverid string,
events bigint,
totmiles bigint,
riskfactor float)
;
```

b) Use the below PIG's script to populate "riskfactor" table:

```
a = LOAD 'geolocation' using org.apache.hive.hcatalog.pig.HCatLoader();
b = filter a by event != 'normal';
c = foreach b generate driverid, event, (int) '1' as occurance;
d = group c by driverid;
e = foreach d generate group as driverid, SUM(c.occurance) as t_occ;
g = LOAD 'drivermileage' using org.apache.hive.hcatalog.pig.HCatLoader();
h = join e by driverid, g by driverid;
final_data = foreach h generate $0 as driverid, $1 as events, $3 as totmiles,
(float)$1/$3*1000000 as riskfactor;
store final_data into 'riskfactor' using org.apache.hive.hcatalog.pig.HCatStorer();
```

8) Create table riskfactor_idling:

Code in PIG:-

```
a = LOAD 'geolocation' using org.apache.hive.hcatalog.pig.HCatLoader();
b = filter a by idling_ind = 1;
c = foreach b generate driverid, event, (int) '1' as occurance;
d = group c by driverid;
e = foreach d generate group as driverid, SUM(c.occurance) as t_occ;
g = LOAD 'drivermileage' using org.apache.hive.hcatalog.pig.HCatLoader();
h = join e by driverid, g by driverid;
final_data = foreach h generate $0 as driverid, $1 as events, $3 as totmiles,
(float)$1/$3*1000000 as riskfactor;
store final_data into 'riskfactor_idling' using org.apache.hive.hcatalog.pig.HCatStorer();
```

Table code:

```
CREATE TABLE riskfactor_idling
(driverid string,
events bigint,
totmiles bigint,
riskfactor float)
;
```

9) Create table riskfactor-3:

```
CREATE TABLE riskfactor_davg_final1
AS
SELECT g.driverid,g.city,g.state, (dm.totmiles/ (avg(g.velocity)*10000)) as risk_driver
FROM geolocation g join drivermileage dm
on g.driverid=dm.driverid

GROUP BY g.driverid,g.city,g.state,dm.totmiles ;
```

10) Create table riskfactor-4:

```
CREATE TABLE riskfactor_mavg_final
AS
SELECT g.truckid,g.city,g.state, (dm.avgmpg/ avg(g.velocity))*10 as risk_truck
FROM geolocation g join avg_mileage dm
on g.truckid=dm.truckid
where g.event="overspeed"
GROUP BY g.truckid,g.city,g.state,dm.avgmpg ;
```