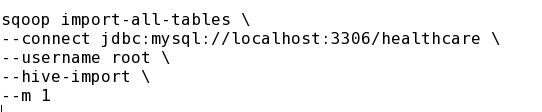
1. Import client data to distributed filesystem of warehouse



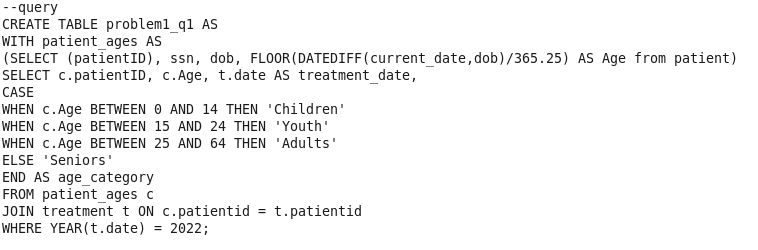


**Question1**: Jimmy, from the healthcare department, has requested a report that shows how the number of treatments each age category of patients \

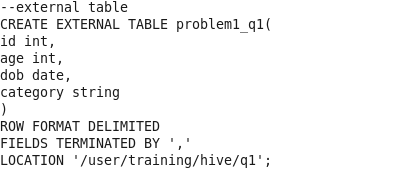
has gone through in the year 2022.

The age category is as follows, Children (00-14 years), Youth (15-24 years), Adults (25-64 years), and Seniors (65 years and over).

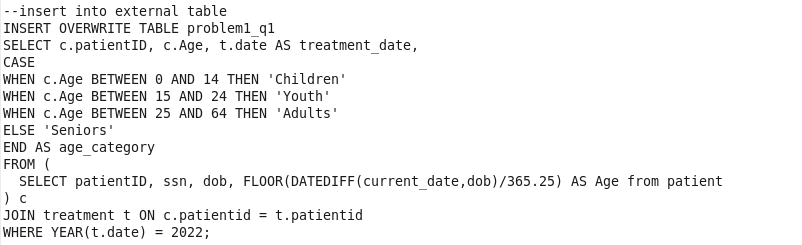
Assist Jimmy in generating the report.

**Hive query:** 

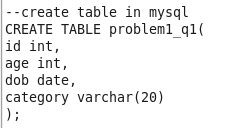
**Create external table**:



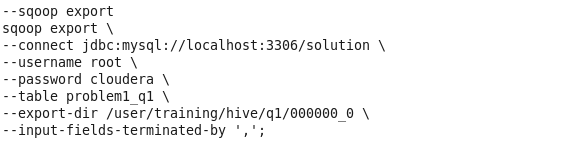
**Insert into external table:**

****

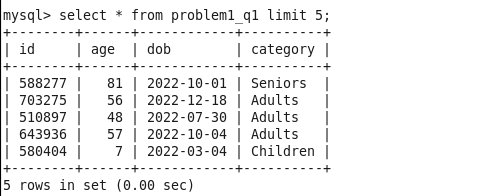
**Create table in client DB:**

****

**Export table using Sqoop:**

****

**Answer in client table:**



**Question2:** Jacob, from insurance management, has noticed that insurance claims are not made for

all the treatments. He also wants to figure out if the gender of the patient has any

impact on the insurance claim. Assist Jacob in this situation by generating a report

that finds for each gender the number of treatments, number of claims, and

treatment-to-claim ratio. And notice if there is a significant difference between

the treatment-to-claim ratio of male and female patients.

--query

SELECT p.gender,

COUNT(t.treatmentId) AS `Total Treatment`,

COUNT(c.claimId) AS `Total Claims`,

COUNT(t.treatmentId) - COUNT(c.claimId) AS `difference`,

COUNT(c.claimId)/COUNT(t.treatmentId) AS `ratio`

FROM treatment t

LEFT JOIN claim c

ON t.claimID=c.claimID

INNER JOIN person p

ON p.personId=t.patientId

GROUP BY p.gender;

--create external table

CREATE EXTERNAL TABLE problem1\_q2(

gender string,

total\_treatment int,

total\_claim int,

difference int,

ratio decimal

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/training/hive/q2';

--insert into external table

INSERT OVERWRITE TABLE problem1\_q2

SELECT p.gender,

COUNT(t.treatmentId) AS `Total Treatment`,

COUNT(c.claimId) AS `Total Claims`,

COUNT(t.treatmentId) - COUNT(c.claimId) AS `difference`,

COUNT(c.claimId)/COUNT(t.treatmentId) AS `ratio`

FROM treatment t

LEFT JOIN claim c

ON t.claimID=c.claimID

INNER JOIN person p

ON p.personId=t.patientId

GROUP BY p.gender;

--create table in mysql

CREATE TABLE problem1\_q2(

gender varchar(20),

total\_treatment int,

total\_claim int,

difference int,

ratio float(10,5)

);

--sqoop export

sqoop export \

--connect jdbc:mysql://localhost:3306/solution \

--username root \

--password cloudera \

--table problem1\_q2 \

--export-dir /user/training/hive/q2/000000\_0 \

--input-fields-terminated-by ',';

**Question3:** The Healthcare department wants a report about the inventory of pharmacies.

Generate a report on their behalf that shows how many units of medicine each

pharmacy has in their inventory, the total maximum retail price of those medicines,

and the total price of all the medicines after discount.

Note: discount field in keep signifies the percentage of discount on the maximum price.

--query

SELECT p.pharmacyID, p.pharmacyName, COUNT(m.medicineid), SUM(m.maxprice), SUM(m.maxprice - (k.discount \* 0.01)) AS `After discount`

FROM pharmacy p

JOIN keep k ON p.pharmacyID = k.pharmacyID

JOIN medicine m ON k.medicineID = m.medicineID

GROUP BY p.pharmacyID, p.pharmacyName;

--create external table

CREATE EXTERNAL TABLE problem1\_q3(

pharmacyid int,

pharmacyname string,

medicine\_count int,

max\_price\_sum float,

discount float

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/training/hive/q3';

-insert in external table

INSERT OVERWRITE TABLE problem1\_q3

SELECT p.pharmacyID, p.pharmacyName, COUNT(m.medicineid), SUM(m.maxprice), SUM(m.maxprice - (k.discount \* 0.01)) AS `After discount`

FROM pharmacy p

JOIN keep k ON p.pharmacyID = k.pharmacyID

JOIN medicine m ON k.medicineID = m.medicineID

GROUP BY p.pharmacyID, p.pharmacyName;

--mysql

CREATE TABLE problem1\_q3(

pharmacyid int,

pharmacyname varchar(20),

medicine\_count int,

max\_price\_sum float,

discount float

);

--sqoop export

sqoop export \

--connect jdbc:mysql://localhost:3306/solution \

--username root \

--password cloudera \

--table problem1\_q3 \

--export-dir /user/training/hive/q3/000000\_0 \

--input-fields-terminated-by ',';

**Question 4 :** Jimmy, from the healthcare department, wants to know which disease is infecting people of which gender more often. Assist Jimmy with this purpose by generating a report that shows for each disease the male-to-female ratio. Sort the data in a way that is helpful for Jimmy.

**--query**

SELECT diseaseid,

SUM(CASE WHEN Gender = 'male' THEN Count END) AS male\_count,

SUM(CASE WHEN Gender = 'female' THEN Count END) AS female\_count,

(SUM(CASE WHEN Gender = 'male' THEN Count END) / SUM(CASE WHEN Gender =

'female' THEN Count END)) AS ratio

FROM (

SELECT d.diseaseID AS diseaseid,

pe.gender AS Gender,

COUNT(\*) AS Count

FROM disease d

JOIN treatment t ON d.diseaseID = t.diseaseID

JOIN patient p ON p.patientID = t.patientID

JOIN person pe ON pe.personID = p.patientID

GROUP BY d.diseaseID, pe.gender

) a

GROUP BY diseaseid

ORDER BY diseaseid ASC;

**--create external table**

CREATE EXTERNAL TABLE problem1\_q4(

id int,

male int,

female int,

ratio decimal(10,5)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/training/hive/q4';

**--insert into external table**

INSERT OVERWRITE TABLE problem1\_q4

SELECT diseaseid,

SUM(CASE WHEN Gender = 'male' THEN Count END) AS male\_count,

SUM(CASE WHEN Gender = 'female' THEN Count END) AS female\_count,

(SUM(CASE WHEN Gender = 'male' THEN Count END) / SUM(CASE WHEN Gender =

'female' THEN Count END)) AS ratio

FROM (

SELECT d.diseaseID AS diseaseid,

pe.gender AS Gender,

COUNT(\*) AS Count

FROM disease d

JOIN treatment t ON d.diseaseID = t.diseaseID

JOIN patient p ON p.patientID = t.patientID

JOIN person pe ON pe.personID = p.patientID

GROUP BY d.diseaseID, pe.gender

) a

GROUP BY diseaseid

ORDER BY diseaseid ASC;

**--mysql create table**

CREATE TABLE problem1\_q4(

id int,

male int,

female int,

ratio decimal(10,5)

);

**--sqoop export**

sqoop export \

--connect jdbc:mysql://localhost:3306/solution \

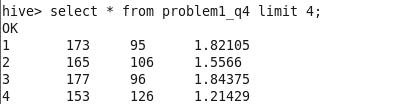
--username root \

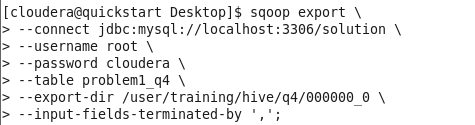
--password cloudera \

--table problem1\_q4 \

--export-dir /user/training/hive/q4/000000\_0 \

--input-fields-terminated-by ',';





**Question 5:** The healthcare department suspects that some pharmacies prescribe more medicines than others in a single prescription, for them, generate a report that finds for each pharmacy the maximum, minimum and the average number of medicines prescribed in their prescriptions.

**--query**

SELECT p.pharmacyName,

MAX(k.quantity) AS max\_meds,

MIN(k.quantity) AS min\_meds,

ROUND(AVG(k.quantity), 2) AS avg\_meds

FROM pharmacy p

JOIN keep k ON p.pharmacyID = k.pharmacyID

GROUP BY p.pharmacyName;

**--create external table**

CREATE EXTERNAL TABLE problem1\_q5

(

pharmacyName STRING,

max\_meds INT,

min\_meds INT,

avg\_meds FLOAT

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/training/hive/q5';

**--insert into external table**

INSERT OVERWRITE TABLE problem1\_q5

SELECT p.pharmacyName,

MAX(k.quantity) AS max\_meds,

MIN(k.quantity) AS min\_meds,

ROUND(AVG(k.quantity), 2) AS avg\_meds

FROM pharmacy p

JOIN keep k ON p.pharmacyID = k.pharmacyID

GROUP BY p.pharmacyName;

**--create table in mysql**

CREATE TABLE problem1\_q5

(

pharmacyName varchar(20),

max\_meds INT,

min\_meds INT,

avg\_meds decimal(10,5)

);

**--sqoop export**

sqoop export \

--connect jdbc:mysql://localhost:3306/solution \

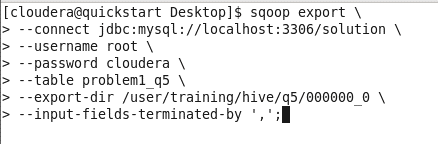
--username root \

--password cloudera \

--table problem1\_q5 \

--export-dir /user/training/hive/q5/000000\_0 \

--input-fields-terminated-by ',';



**Question 6:** A company needs to set up 3 new pharmacies, they have come up with an idea that the pharmacy can be set up in cities where the pharmacy-to-prescription ratio is the lowest and the number of prescriptions should exceed 100. Assist the company to identify those cities where the pharmacy can be set up.

**--query**

SELECT a.city,

ROUND(COUNT(DISTINCT pha.pharmacyID) / COUNT(DISTINCT pre.prescriptionID),4)

AS pharmacy\_prescription\_ratio

FROM pharmacy pha

JOIN address a ON a.addressID = pha.addressID

JOIN prescription pre ON pha.pharmacyID = pre.pharmacyID

GROUP BY a.city

HAVING COUNT(pre.prescriptionID) > 100

ORDER BY pharmacy\_prescription\_ratio;

**--create internal table**

CREATE EXTERNAL TABLE problem1\_q6

(

city string,

ratio decimal(6,4)

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/training/hive/q6';

**--insert into external table**

INSERT OVERWRITE TABLE problem1\_q6

SELECT a.city,

ROUND(COUNT(DISTINCT pha.pharmacyID) / COUNT(DISTINCT pre.prescriptionID),4)

AS pharmacy\_prescription\_ratio

FROM pharmacy pha

JOIN address a ON a.addressID = pha.addressID

JOIN prescription pre ON pha.pharmacyID = pre.pharmacyID

GROUP BY a.city

HAVING COUNT(pre.prescriptionID) > 100

ORDER BY pharmacy\_prescription\_ratio;

**--create table mysql**

CREATE TABLE problem1\_q6

(

city varchar(10),

ratio decimal(6,4)

);

**--sqoop export**

sqoop export \

--connect jdbc:mysql://localhost:3306/solution \

--username root \

--password cloudera \

--table problem1\_q6 \

--export-dir /user/training/hive/q6/000000\_0 \

--input-fields-terminated-by ',';



**Question 7:** The State of Alabama (AL) is trying to manage its healthcare resources more efficiently. For each city in their state, they need to identify the disease for which the maximum number of patients have gone for treatment. Assist the state for this purpose. Note: The state of Alabama is represented as AL in Address Table.

**--bucketing**

CREATE TABLE address\_part (

addressid INT,

address1 STRING,

city STRING,

zip INT)

PARTITIONED BY (state STRING) CLUSTERED BY (city) INTO 10 BUCKETS;

INSERT INTO TABLE address\_part PARTITION (state)SELECT addressid,address1,city,zip,state FROM address;

**--CREATE EXTERNAL TABLE**

CREATE EXTERNAL TABLE prb1\_7 (city STRING, diseaseName STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

**-insert into external table**

INSERT OVERWRITE TABLE prb1\_7 SELECT city, diseaseName

FROM (

SELECT city, diseaseName, counts, DENSE\_RANK() OVER (PARTITION BY city ORDER BY

counts DESC) AS ranks

FROM (

SELECT a.city, d.diseaseName, COUNT(t.patientID) AS counts

FROM treatment t

JOIN person p ON t.patientID = p.personID

JOIN disease d ON t.diseaseID = d.diseaseID

JOIN address\_part a ON p.addressID = a.addressID

WHERE a.state = 'AL'

GROUP BY a.city, d.diseaseName

) a

) b

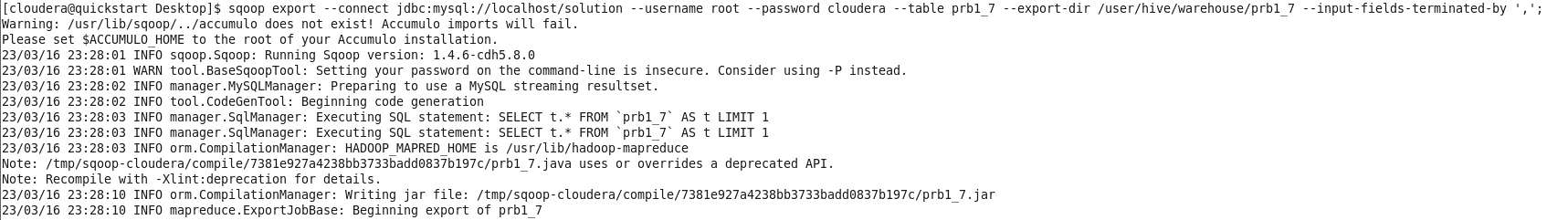
WHERE ranks = 1;

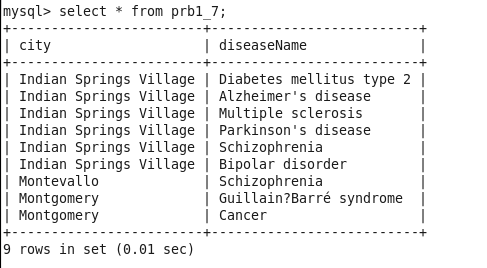
**--mysql table**

create table prb1\_7 (city VARCHAR(30), diseaseName VARCHAR(50)) ;

**--sqoop export**

sqoop export --connect jdbc:mysql://localhost/solution --username root --password cloudera --table prb1\_7 --export-dir /user/hive/warehouse/prb1\_7 --input-fields-terminated-by ',';





**Question 8:** The Healthcare department wants to know which disease is most likely to infect multiple people in the same household. For each disease find the number of households that has more than one patient with the same disease. Note: 2 people are considered to be in the same household if they have the same address.

**--BUCKETING**

CREATE TABLE treatment\_part (

treatmentid INT,

date DATE,

patientid INT,

diseaseid INT,

claimid BIGINT

)

CLUSTERED BY (diseaseid) INTO 5 BUCKETS;

INSERT INTO TABLE treatment\_part SELECT treatmentid,date,patientid,diseaseid,claimid

FROM treatment

**--external table**

CREATE EXTERNAL TABLE prb1\_8 (diseaseName STRING, no\_of\_household INT) ROW

FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

**--insert into external table**

INSERT OVERWRITE TABLE prb1\_8 SELECT diseaseName, COUNT(address1) AS

no\_of\_household

FROM (select d.diseaseName,a.address1,COUNT(t.patientID)

from disease d join treatment\_part t on d.diseaseID=t.diseaseID

join person p on t.patientID=p.personID

join address a on p.addressID=a.addressID

group by d.diseasename,a.address1

having count(t.patientID)>1

order by d.diseasename desc,a.address1) a

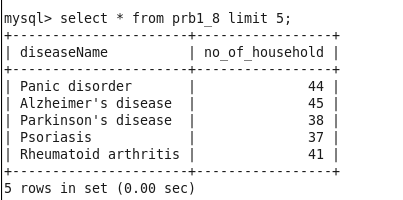
group by diseaseName;

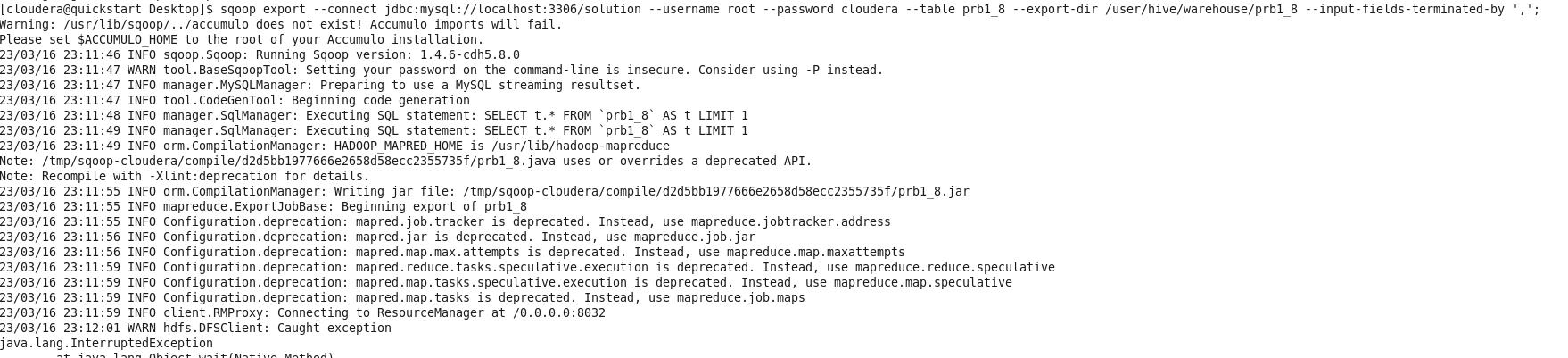
**--mysql table**

create table prb1\_8 (diseaseName VARCHAR(50), no\_of\_household INT);

**--sqoop export**

sqoop export --connect jdbc:mysql://localhost:3306/solution --username root --password cloudera --table prb1\_8 --export-dir /user/hive/warehouse/prb1\_8 --input-fields-terminated-by ',';





**Question 9:** An Insurance company wants a state wise report of the treatments to claim ratio between 1st April 2021 and 31st March 2022 (days both included). Assist them to create such a report.

**--query**

select a.state, round(count(treatmentID)/sum(if(claimID is

null,0,1)),2) as treatment\_claim\_ratio

from treatment\_part t join person p on t.patientID=p.personID

join address\_part a on a.addressID=p.addressID

where t.date between '2021-04-01' and '2022-03-31'

group by a.state;

**--external table**

CREATE EXTERNAL TABLE prb1\_9 (state STRING, treatment\_claim\_ratio DECIMAL(3,2)) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

**-- insert into external table**

INSERT OVERWRITE TABLE prb1\_9 select a.state, round(count(treatmentID)/sum(if(claimID is

null,0,1)),2) as treatment\_claim\_ratio

from treatment\_part t join person p on t.patientID=p.personID

join address\_part a on a.addressID=p.addressID

where t.date between '2021-04-01' and '2022-03-31'

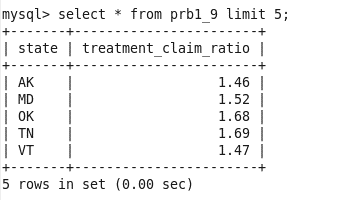
group by a.state;

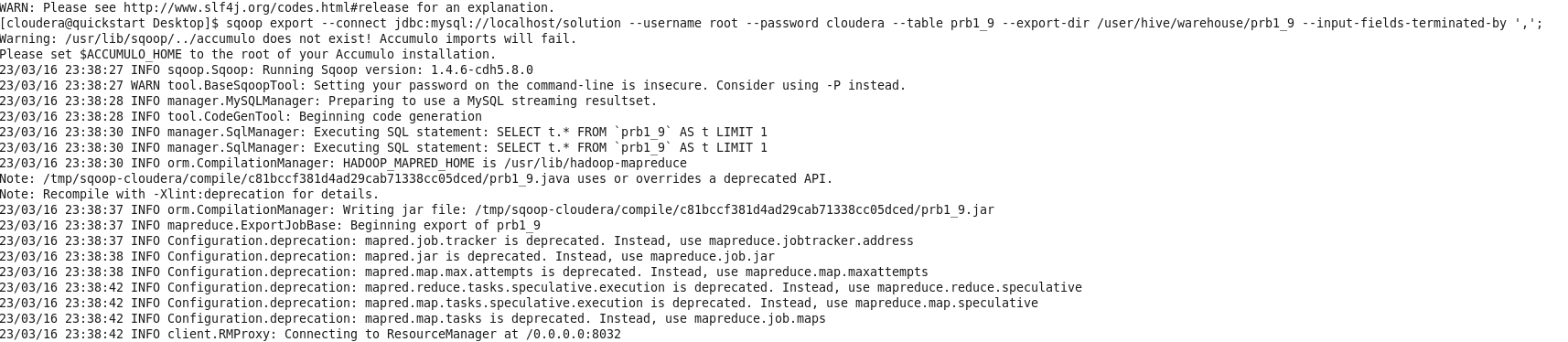
**--mysql table**

create table prb1\_9 (state VARCHAR(2), treatment\_claim\_ratio DECIMAL(3,2));

**--sqoop export**

sqoop export --connect jdbc:mysql://localhost/solution --username root --password cloudera --table prb1\_9 --export-dir /user/hive/warehouse/prb1\_9 --input-fields-terminated-by ',';



****