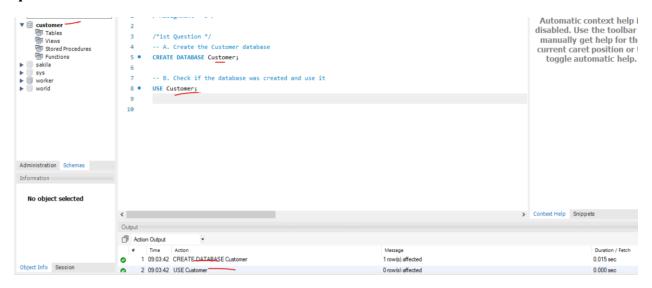
INFO 531: DATA WAREHOUSING IN THE CLOUD

ASSIGNMENT-3

Q1. $\{A\}$ Using the MySQL Workbench, create a database called Customer. The database must be named "Customer". $\{B\}$ Check if the database was created and use the same for further questions.



Q2. { A } Create a staging table, ** Customer.CustomerChurn_Stage **, in a database system, with the column list provided in the CSV file. Define the 'CustomerId' as the Primary Key (PK). Get the table definition (DDL) from the database system and capture it in a Word document for submission. { B } Create a persistent table, ** Customer.CustomerChurn **, with the column list provided in the CSV file + following 5 columns: << SourceSystemNm NVARCHAR(20) NOT NULL, CreateAgentId NVARCHAR(20) NOT NULL, CreateDtm DATETIME NOT NULL, ChangeAgentId NVARCHAR(20) NOT NULL, ChangeDtm DATETIME NOT NULL >> Define the 'CustomerId' as the Primary Key (PK). Get the table definition (DDL) from the database system and capture it in a Word document for submission.

```
-- A. Create staging table CustomerChurn_Stage

CREATE TABLE Customer.CustomerChurn_Stage (

CustomerId INT,

Surname VARCHAR(255),

CreditScore INT,

Geography VARCHAR(255),

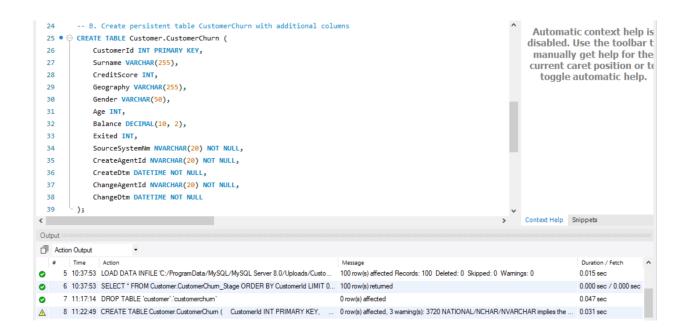
Gender VARCHAR(50),

Age INT,

Balance DECIMAL(10, 2),

Exited INT

);
```



Q3. { A } Load the staging table, ** Customer.CustomerChurn_Stage **, with data from the CSV file, CustomerChurn1.csv . { B } Verify data by comparing the row counts between the CSV file and the staging table, ** Customer.CustomerChurn_Stage [Data Source: CustomerChurn1.CSV] **. Provide the screenshot of last few rows using the 'SELECT *'. Make sure the output shows all column values. The SELECT statement must use the ORDER BY 'CustomerId'.

```
-- A. Load data into CustomerChurn_Stage from CustomerChurn1.csv

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/CustomerChurn1.csv'

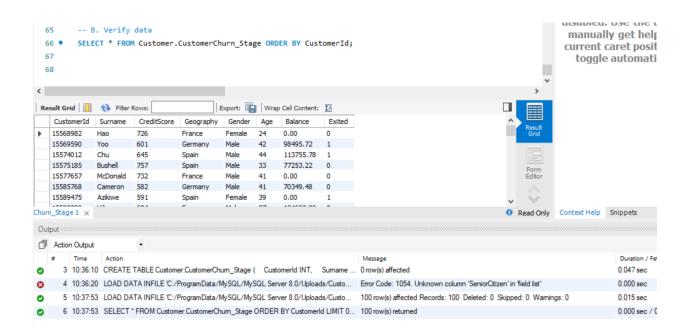
INTO TABLE Customer.CustomerChurn_Stage

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 ROWS

(CustomerId, Surname, CreditScore, Geography, Gender, Age, Balance, Exited);
```

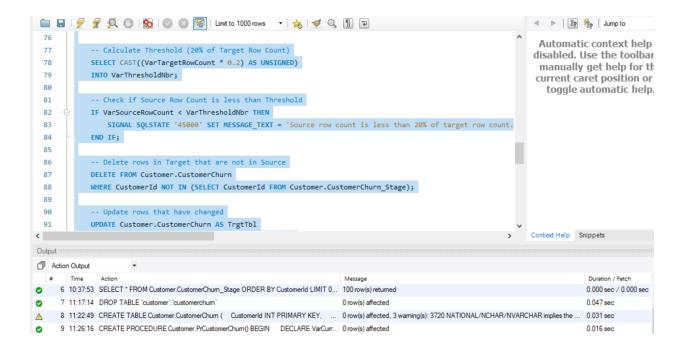


Q4. Create a database stored procedure based on the template provided along with this assignment << StoredProc_Template.txt >>. Name the stored procedure name this: ** Customer.PrCustomerChurn ** . [[NOTE : This stored procedure will use the table, ** Customer.CustomerChurn_Stage ** , as the source (aka, staging table). This stored procedure will use the table, ** Customer.CustomerChurn **, as the target (aka, persistent table).]]

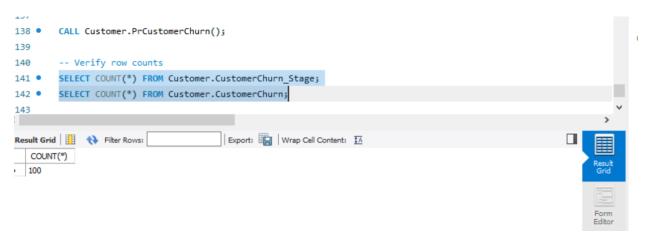
```
SQL qUERY:
      DELIMITER //
CREATE PROCEDURE Customer.PrCustomerChurn()
BEGIN
 DECLARE VarCurrentTimestamp TIMESTAMP DEFAULT CURRENT_TIMESTAMP;
 DECLARE VarSourceRowCount, VarTargetRowCount, VarThresholdNbr INTEGER DEFAULT 0;
 DECLARE VarTinyIntVal TINYINT;
 -- Get Source and Target Row Counts
 SELECT COUNT(*)
 INTO VarSourceRowCount
 FROM Customer.CustomerChurn_Stage;
 SELECT COUNT(*)
 INTO VarTargetRowCount
 FROM Customer.CustomerChurn:
 -- Calculate Threshold (20% of Target Row Count)
 SELECT CAST((VarTargetRowCount * 0.2) AS UNSIGNED)
 INTO VarThresholdNbr;
```

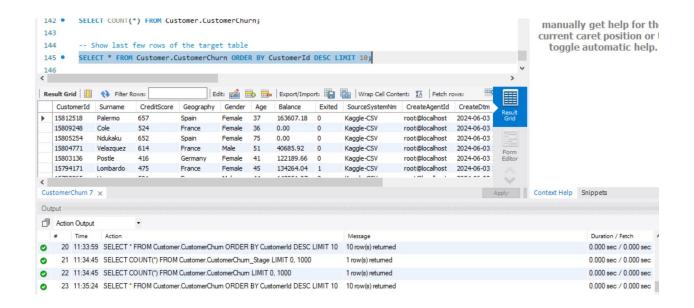
```
-- Check if Source Row Count is less than Threshold
 IF VarSourceRowCount < VarThresholdNbr THEN
   SIGNAL SQLSTATE '45000' SET MESSAGE_TEXT = 'Source row count is less than 20% of target
row count.';
 END IF;
 -- Delete rows in Target that are not in Source
 DELETE FROM Customer.CustomerChurn
 WHERE CustomerId NOT IN (SELECT CustomerId FROM Customer.CustomerChurn Stage);
 -- Update rows that have changed
 UPDATE Customer.CustomerChurn AS TrgtTbl
 INNER JOIN Customer.CustomerChurn_Stage AS SrcTbl
  ON TrgtTbl.CustomerId = SrcTbl.CustomerId
 SET TrgtTbl.Surname = SrcTbl.Surname,
   TrgtTbl.CreditScore = SrcTbl.CreditScore,
   TrgtTbl.Geography = SrcTbl.Geography,
   TrgtTbl.Gender = SrcTbl.Gender,
   TrgtTbl.Age = SrcTbl.Age,
   TrgtTbl.Balance = SrcTbl.Balance,
   TrgtTbl.Exited = SrcTbl.Exited,
   TrgtTbl.ChangeAgentId = CURRENT_USER(),
   TrgtTbl.ChangeDtm = VarCurrentTimestamp
 WHERE (
   COALESCE(TrgtTbl.Surname, '*') <> COALESCE(SrcTbl.Surname, '*') OR
```

```
COALESCE(TrgtTbl.CreditScore, '*') <> COALESCE(SrcTbl.CreditScore, '*') OR
   COALESCE(TrgtTbl.Geography, '*') <> COALESCE(SrcTbl.Geography, '*') OR
   COALESCE(TrgtTbl.Gender, '*') <> COALESCE(SrcTbl.Gender, '*') OR
   COALESCE(TrgtTbl.Age, '*') <> COALESCE(SrcTbl.Age, '*') OR
   COALESCE(TrgtTbl.Balance, '*') <> COALESCE(SrcTbl.Balance, '*') OR
   COALESCE(TrgtTbl.Exited, '*') <> COALESCE(SrcTbl.Exited, '*')
 );
 -- Insert new rows
 INSERT INTO Customer.CustomerChurn (
   CustomerId, Surname, CreditScore, Geography, Gender, Age, Balance, Exited,
   SourceSystemNm, CreateAgentId, CreateDtm, ChangeAgentId, ChangeDtm
 )
 SELECT
   SrcTbl.CustomerId, SrcTbl.Surname, SrcTbl.CreditScore, SrcTbl.Geography,
   SrcTbl.Gender, SrcTbl.Age, SrcTbl.Balance, SrcTbl.Exited,
   'Kaggle-CSV', CURRENT_USER(), VarCurrentTimestamp, CURRENT_USER(),
VarCurrentTimestamp
 FROM Customer.CustomerChurn_Stage AS SrcTbl
 LEFT JOIN Customer.CustomerChurn AS TrgtTbl
 ON SrcTbl.CustomerId = TrgtTbl.CustomerId
 WHERE TrgtTbl.CustomerId IS NULL;
END //
DELIMITER;
```



Q5. Execute the stored procedure, ** Customer.PrCustomerChurn **, that was created in Q4. After execution, the stored procedure should load data from the stage to the persistent table: ** Customer.CustomerChurn **. {A} Verify data by comparing the row counts between the staging table, ** Customer.CustomerChurn_Stage [Data Source: CustomerChurn1.CSV] ** and the persistent table: ** Customer.CustomerChurn **. { B } Provide the screenshot of last few rows using the SELECT *. Make sure the output shows all column values. The SELECT statement must use the ORDER BY CustomerId.





Q6. After data verification is completed, in Q5 , { A } create table, **

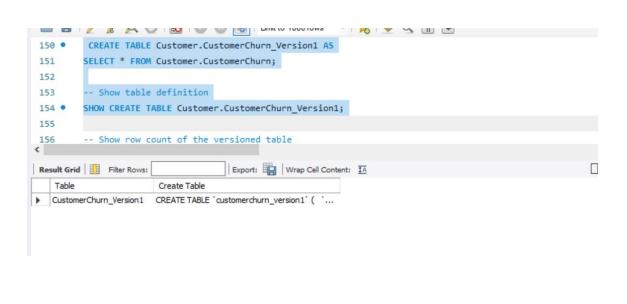
Customer.CustomerChurn_Version1 **, with data from ** Customer.CustomerChurn ** (that was already loaded from Customer.CustomerChurn_Stage via the stored procedure). { B } Show table definition of Customer.CustomerChurn_Version1 and show the row count of the table, ** Customer.CustomerChurn_Version1 **: { C } Provide the screenshot of last few rows for ** Customer.CustomerChurn_Version1 ** [Originally data came from:

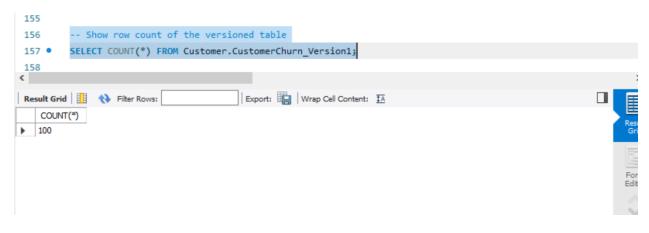
CustomerChurn1.CSV]. Make sure the output shows all column values. The SELECT statement must use the ORDER BY CustomerId. { D } Empty the staging table, **

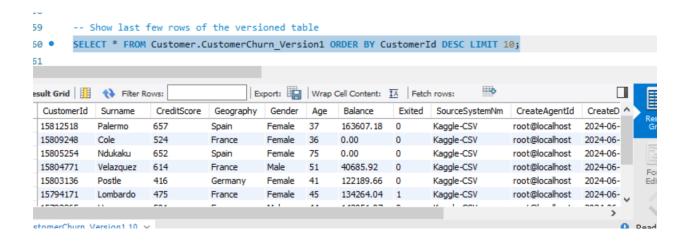
Customer.CustomerChurn_Stage **, and load it with data from the CSV file,

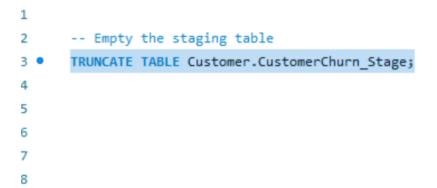
"CustomerChurn2.csv". Verify data by comparing the row counts between the CSV file and

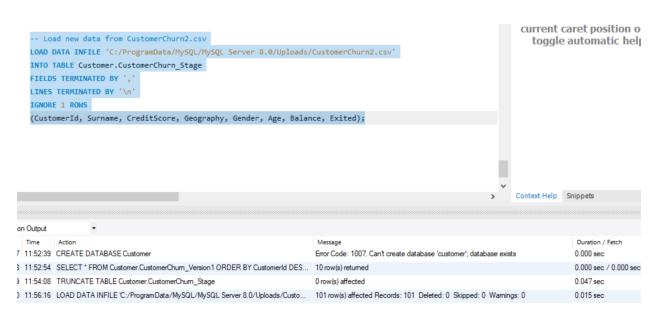
the staging table, ** Customer.CustomerChurn_Stage ** [Data Source: CustomerChurn2.CSV]. Provide the row count of ** Customer.CustomerChurn_Stage ** that you loaded from CustomerChurn2.csv file. Provide the screenshot of last few rows using the SELECT *. Make sure the output shows all column values. The SELECT statement must use the ORDER BY CustomerId.

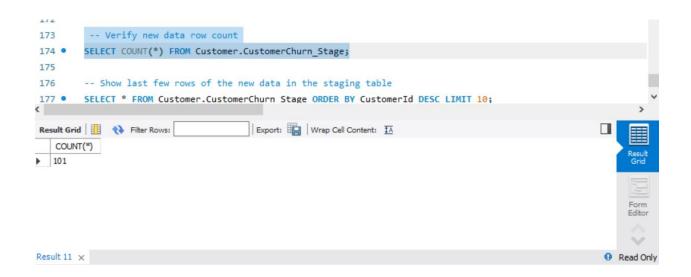


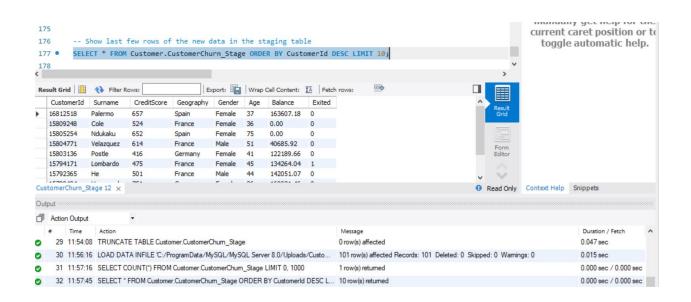




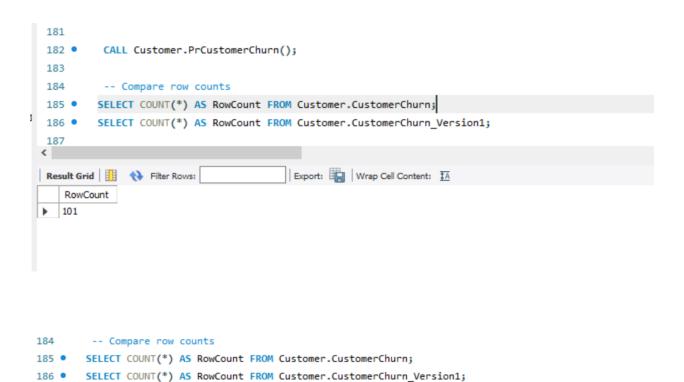








Q7. Execute the stored procedure, Customer.PrCustomerChurn, that was created in Q4. After execution, the stored procedure should load data from the stage to the persistent table: Customer.CustomerChurn. CALL `customer`.`PrCustomerChurn`(); This time, the table will be refreshed via DELETE, UPDATE, and INSERT/SELECT statements in the stored procedure. Show the row count results of both Customer.CustomerChurn_Version1 table [Data Source: CustomerChurn1.CSV] and the persistent table: Customer.CustomerChurn. Compare the rows between the Customer.CustomerChurn_Version1 [Data Source: CustomerChurn1.CSV] table and the persistent table: Customer.CustomerChurn [Data Source: CustomerChurn2.CSV]. Show the rows that are available in the Customer.CustomerChurn_Version1 table but not in the Customer.CustomerChurn table (implementation of brand-new row DELETE statement of the stored procedure).



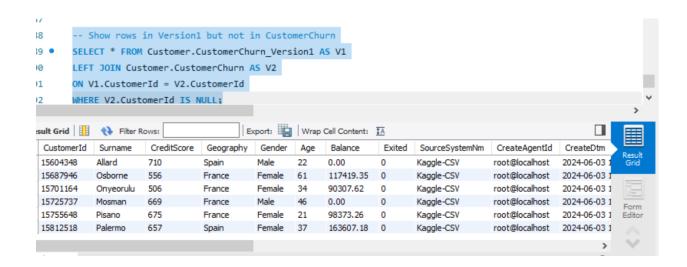
Export: Wrap Cell Content: 1A

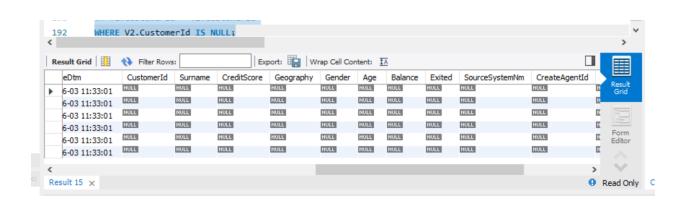
187

▶ 100

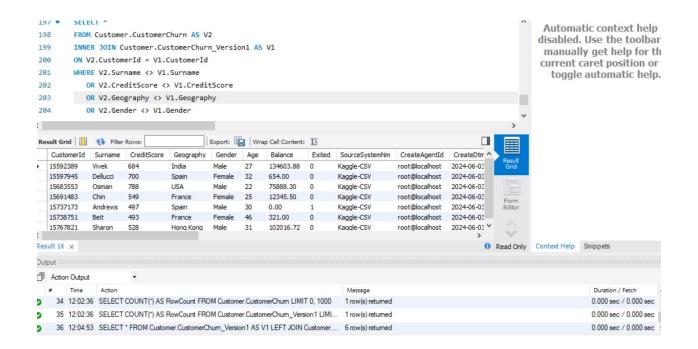
Result Grid Filter Rows:

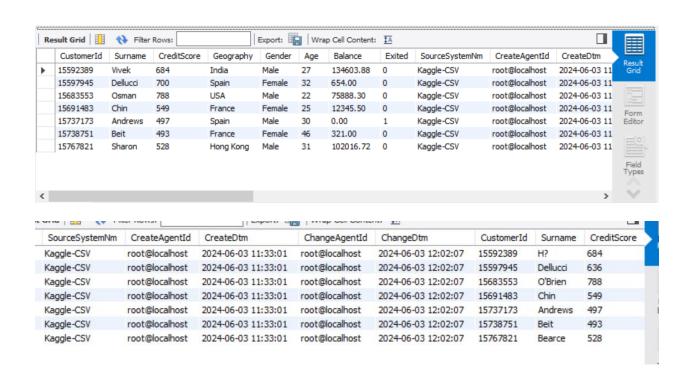
RowCount

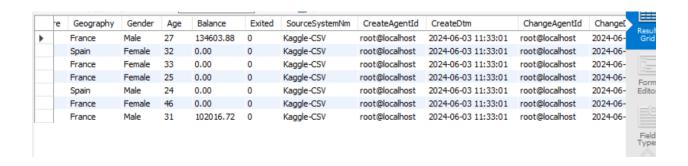




Q8. Show the rows (SELECT *) that changed (one or many non-Primary Key columns), in the Customer.CustomerChurn table (implementation of UPDATE statement of the stored procedure). You need to perform a comparison between Customer.CustomerChurn table [Data Source: CustomerChurn2.CSV] and Customer.CustomerChurn_Version1 table [Data Source: CustomerChurn1.CSV] in terms of non-PK columns (Excluds: SourceSystemNm, CreateAgentId, CreateDtm, ChangeAgentId, ChangeDtm), and with a join condition using the PK column(s). You must do ORDER BY CustomerId. The output of this query should show different values for the CreateDtm and ChangeDtm columns in Customer.CustomerChurn table for the changed rows. Take a screenshot and capture it in the Word document. Make sure all columns including CreateDtm and ChangeDtm of CustomerChurn table are displayed.







Q9. Provide the screenshot of last few rows using the SELECT * FROM Customer.CustomerChurn. Make sure the output shows all column values. The SELECT statement must use the ORDER BY CustomerId. Show the rows that are available in the Customer.CustomerChurn table [Data Source: CustomerChurn2.CSV] but not in the Customer.CustomerChurn_Version1 table (implementation of brand-new rows INSERT by the stored procedure). Do a SELECT * along with ORDER BY CustomerId. Take a screenshot and capture it in the Word document.

