1. Analysis: Advanced analysis of Sample Superstore spreadsheet. Identify and document entities, attributes, domains, referential integrity

Sample – Superstore spreadsheet consists of 9994 rows and 21 columns. 6 of these columns are numerical, 2 are date columns and the rest 13 columns are categorical. Number of unique values for these columns are as follows:

```
df.nunique()

Row ID            9994
Order ID          5009
Order Date        1236
Ship Date         1334
Ship Mode            4
Customer ID        793
Customer Name      793
Segment              3
Country/Region       1
City               531
State               49
Postal Code        630
Region               4
Product ID        1862
Category             3
Sub-Category        17
Product Name      1849
Sales             6144
Quantity            14
Discount            12
Profit            7545
```

- As you can see, Order ID is not unique for every row in our spreadsheet. It happens because an order can consist of different products and the products' price related information on that order are stored on different rows.
- There are 4 shipping modes on the spreadsheet: Standard Class, Second Class, First Class and Same Day.
- Number of unique Customer ID is the same as the number of unique Customer Name with the value of 793. This is because every customer is assigned only one id.
- Segment column has 3 unique values: Consumer, Corporate and Home Office
- The only country in this spreadsheet is United States and has at least one order for every US state except Alaska and Hawaii.
- Postal Code column is the only column with NaN values. We discovered that all NaN Postal Codes belong to the Burlington city of the Vermont state. Thus we can fill these Nan values with 05401 which is a postal code for Burlington, Vermont with the highest population.
- There are also 438 Postal Code values with a missing 0 digit at the start. We can simply fix it by adding a 0 at the beginning of those values.
- West, East, Central and South are the values for Region column.
- Office Supplies, Furniture and Technology are the values for Category column.
- There are 17 Sub-Categories in the spreadsheet.
    - Office Supplies has 9 subcategories
    - Furniture has 4 subcategories
    - And Technology has 4 subcategories

When converting this spreadsheet into a database, storing every value on a single table would raise problems. Since some columns are related to each other and others are related to different columns we have to split our spreadsheet into various tables by considering referential integrity.

- There are different ID values and feature columns for customers, orders, products, regions and addresses.
    - We should create a customer table with CustomerID as a primary key and customer related columns as features.
    - Region table should have information about RegionCode and RegionName and this table should have a one-to-many relation with the address table.
    - Address table should be created by using PostalCode as a primary key and RegionCode as a foreign key.
    - Order Table should take foreign keys from customer and address tables as CustomerID, PostalCode. Since every row in this table gives information about a single order, OrderID is the primary key of this table.
    - Information about products should be stored in the Product table with ProductID as the primary key.
    - And finally, Sales master table should be formed with one-to-many relation from the Product and Order tables. ProductID and OrderID combinations will form the primary keys in this table.

2. Logical-level ERD: create an ERD indicating entities, attributes, relationships, primary and foreign keys, and Master Data tables. Include a Metadata table, an optionally a Reference Data table.

Note: you can create new entities and attributes if needed. Use standard notation (crowfoot) and any tools such as Lucidchart, Visio, SQL Workbench

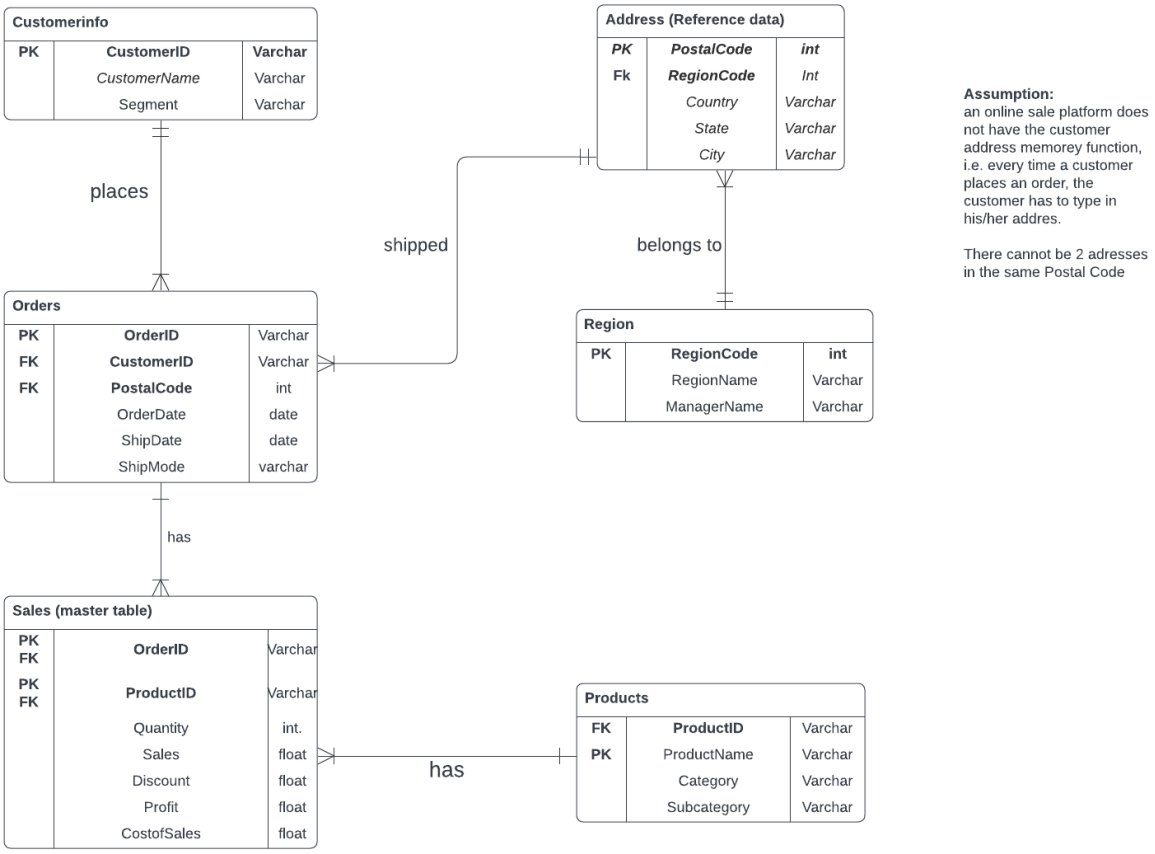


Fig1: Logical-level ERD for Superstore Dataset

Meta-data tables:

Customer - Meta data		
Attribute	**Data type**	**Description**
Customer_id	varchar(255)	Primary key of the table that uniquely idenitfies customer
CustomerName	varchar(255)	Name of a customer
Segment	varchar(255)	Type of customer - (Consumer, Corporate, HomeOffice)

Order - Meta data		
Attribute	**Data type**	**Description**
Order_id	varchar(255)	Primary key of the table that uniquely identifies order
Customer_id	varchar(255)	Foreign key of the table that identifies which customer the order belongs to
PostalCode	int(5)	PostalCode where the order is being shipped
OrderDate	date	Date when order was taken
ShipDate	date	Date when the order was shipped
ShipMode	varchar(255)	Mode of shipping - (StandardClass, FirstClass, SecondClass)

Address - Meta data		
Attribute	**Data type**	**Description**
PostalCode	int(5)	Primary key of the table that uniquely idenitfies postalCode
RegionCode	int(20)	Foreign key that idenifies which region the postalcode belongs to
Country	varchar(255)	Name of the country
State	varchar(255)	Name of the State
City	varchar(255)	Name of the City

Region - Meta data		
Attribute	**Data type**	**Description**
RegionCode	int(5)	Primary key uniquely idenifies regions
RegionName	varchar(255)	East, West, Central, South
ManagerName	varchar(255)	Name of the regional manager

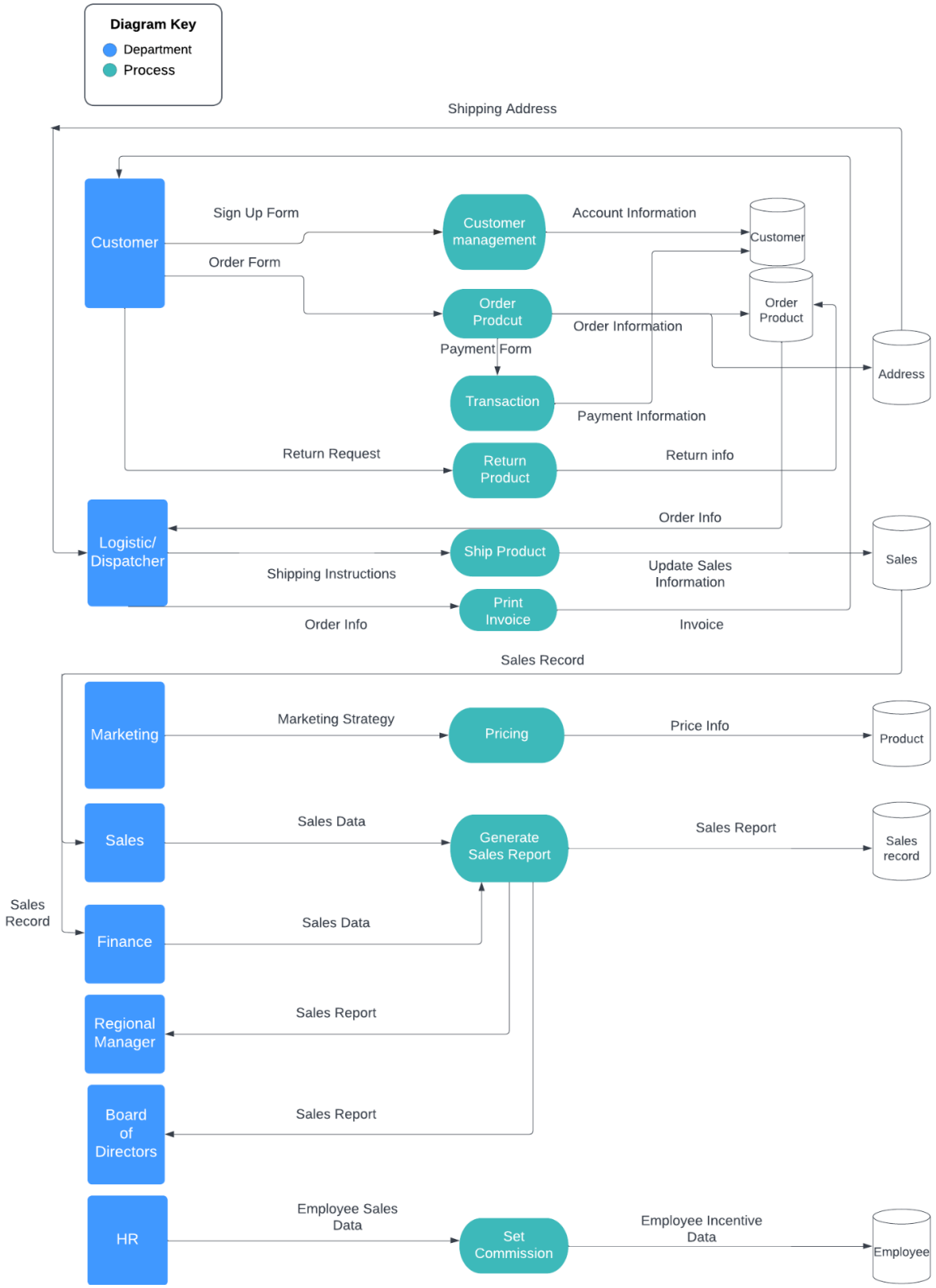
Sales - Meta data		
Attribute	**Data type**	**Description**
Order_id	varchar(255)	One of the composite keys that uniquely identifies sales of a product belonging to specific order
Product_id	varchar(255)	Second composite key that uniquely identifies sales for specific product belonging to an order
Quantity	int(20)	Number of products being ordered/shipped
Sales	float(20)	Total sales of a product that has been ordered
Discount	float(20)	Discount on purchase of a product
Profit	float(20)	Total Profit
CostofSales	float(20)	Total cost of sales i.e. Sales*quantiy - profit

Product - Meta data		
Attribute	**Data type**	**Description**
Product_id	varchar(255)	Primary key of the table that uniquely idenitfies every product
ProductName	varchar(255)	Name of the product
Category	varchar(255)	Furniture, OfficeSupplies, Technology
Sub-Category	varchar(255)	17 sub categories belonging to the Category

Reference Data:

The address table is the reference data for the entire dataset. It consists of 5 columns and 633 rows. For every region, state and city of a country, this data table acts as a reference. This data is obtained from a free source https://www.unitedstateszipcodes.org/.

3. Data Flow: create a data flow indicating data in/out to/from the corresponding processes and the connections among them. Use any tools such as Visio, Lucidchart



Data Flow Diagram

4. Data Cleansing

1. Null postal code for Vermont

Country	City	State	Postal C
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	
United Stat	Burlington	Vermont	

FILLING: replacing with local postal code

(as we search online)

ZIP Code 05401	Burlington	Chittenden	Standard
ZIP Code 05402	Burlington	Chittenden	P.O. Box
ZIP Code 05403	South Burlington	Chittenden	Standard
ZIP Code 05404	Winooski	Chittenden	Standard
ZIP Code 05405	Burlington	Chittenden	Unique
ZIP Code 05406	Burlington	Chittenden	P.O. Box
ZIP Code 05407	South Burlington	Chittenden	P.O. Box
ZIP Code 05408	Burlington	Chittenden	Standard

Some northeastern cities have five-digit postal codes which should have started with 0; Part of them did start with 0 (5 digits), and the rest didn't start with 0 (4 digits) in the data. Here, we choose to start with 0:

A	B	C	D	E	F	G	H	I	J	K	L	M
Row ID	Order ID	Order Date	Ship Date	Ship Mo	Custom	Custom	Segmer	Country	City	State	Postal C	Region
2235	CA-2021-104066	12/5/2021	12/10/2021	Standard (	QJ-19255	Quincy Jor	Corporate	United Stat	Burlington	Vermont	05401	East
5275	CA-2019-162887	11/7/2019	11/9/2019	Second Cl	SV-20785	Stewart Vi	Consumer	United Stat	Burlington	Vermont	05402	East
8799	US-2020-150140	4/6/2020	4/10/2020	Standard (	VM-21685	Valerie Mit	Home Offic	United Stat	Burlington	Vermont	05405	East
9147	US-2020-165505	1/23/2020	1/27/2020	Standard (	CB-12535	Claudia Be	Corporate	United Stat	Burlington	Vermont	05406	East
9148	US-2020-165505	1/23/2020	1/27/2020	Standard (	CB-12535	Claudia Be	Corporate	United Stat	Burlington	Vermont	05406	East
9149	US-2020-165505	1/23/2020	1/27/2020	Standard (	CB-12535	Claudia Be	Corporate	United Stat	Burlington	Vermont	05406	East
9387	US-2021-127292	1/19/2021	1/23/2021	Standard (	RM-19375	Raymond I	Consumer	United Stat	Burlington	Vermont	05408	East
9388	US-2021-127292	1/19/2021	1/23/2021	Standard (	RM-19375	Raymond I	Consumer	United Stat	Burlington	Vermont	05408	East
9389	US-2021-127292	1/19/2021	1/23/2021	Standard (	RM-19375	Raymond I	Consumer	United Stat	Burlington	Vermont	05408	East
9390	US-2021-127292	1/19/2021	1/23/2021	Standard (	RM-19375	Raymond I	Consumer	United Stat	Burlington	Vermont	05408	East
9742	CA-2019-117086	11/8/2019	11/12/2019	Standard (	QJ-19255	Quincy Jor	Corporate	United Stat	Burlington	Vermont	05401	East

2. duplicates (Row 3406 and Row 3407 are the same)

A	B	C	F	N	R	S
Row ID	Order ID	Order Date	Custom	Product ID	Sales	Quantity
3406	US-2018-150119	4/23/2018	LB-16795	FUR-CH-10002965	281.372	2
3407	US-2018-150119	4/23/2018	LB-16795	FUR-CH-10002965	281.372	2

DELETION: deleting Row 3407

Final:

	A	B	C	F	N	R	S
	Row ID	Order ID	Order Date	Custom	Product ID	Sales	Quantity
7	3406	US-2018-150119	4/23/2018	LB-16795	FUR-CH-10002965	281.372	2
5							

Also adjust the following Row ID.

3405	CA-2019-152527
3406	US-2018-150119
3407	US-2018-150119
3408	US-2018-150119
3409	US-2021-150847

3. Product name with multiple ID & Product ID with 2 Product name

UPDATE

a. There are 17 Product name having more than 1 Product ID,

Product Name	Category	Sub-Categ	Total
#10- 4 1/8" x 9 1/2" Recycled Envelopes	Office Supplies	Envelopes	2
Avery Non-Stick Binders	Office Supplies	Binders	2
DAX Wood Document Frame	Furniture		2
Easy-staple paper	Office Supplies	Paper	8
Eldon Wave Desk Accessories	Furniture		2
KI Adjustable-Height Table	Furniture		2
Okidata C610n Printer	Technology	Machines	2
Peel & Seel Recycled Catalog Envelopes, Brown	Office Supplies	Envelopes	2
Prang Drawing Pencil Set	Office Supplies	Art	2
Staple envelope	Office Supplies	Envelopes	9
Staple holder	Office Supplies	Appliances	3
Staple magnet	Office Supplies	Storage	2
Staple remover	Office Supplies	Supplies	3
Staple-based wall hangings	Furniture		2
Staples	Office Supplies	Fasteners	10
Staples in misc. colors	Office Supplies	Art	7
Storex Dura Pro Binders	Office Supplies	Binders	2

We randomly assign one of the Product IDs to each Product Name

b. There are 32 Product ID with 2 Product Name

Product ID	Count of Product Name
FUR-BO-10002213	2
FUR-CH-10001146	2
FUR-FU-10001473	2
FUR-FU-10004017	2
FUR-FU-10004091	2
FUR-FU-10004270	2
FUR-FU-10004848	2
FUR-FU-10004864	2
OFF-AP-10000576	2
OFF-AR-10001149	2
OFF-BI-10002026	2
OFF-BI-10004632	2
OFF-BI-10004654	2
OFF-PA-10000357	2
OFF-PA-10000477	2
OFF-PA-10000659	2
OFF-PA-10001166	2
OFF-PA-10001970	2
OFF-PA-10002195	2
OFF-PA-10002377	2
OFF-PA-10003022	2
OFF-ST-10001228	2
OFF-ST-10004950	2
TEC-AC-10002049	2
TEC-AC-10002550	2
TEC-AC-10003832	2
TEC-MA-10001148	2
TEC-PH-10001530	2
TEC-PH-10001795	2
TEC-PH-10002200	2
TEC-PH-10002310	2
TEC-PH-10004531	2
Grand Total	64

We create a new Product ID for the second Product Name within each Product ID. The rule of the new Product ID is (first 3 letters of Category) - (first 2 letters of subcate) - 8 numbers.

- **Entity**

1.Combine excel sheets

a. Use vlookup to combine “People” and “Orders”

M	V
Region	RegionManager
South	Fred Suzuki
South	Fred Suzuki
West	Sadie Pawthorne
South	Fred Suzuki

(columns between these two are hidden)

b. Under “returns”, check if these listed orders are fully returned, by pivot table and vlookup.

"Orders"			"Returns"				
Order ID	Total		Order ID	Total		lookup value	equal or not
CA-2018-100006	1		CA-2018-100762	4		4	yes
CA-2018-100090	2		CA-2018-100867	1		1	yes
CA-2018-100293	1		CA-2018-102652	4		4	yes
CA-2018-100328	1		CA-2018-103373	1		1	yes

The answer is yes. All products under listed orders are returned.

Come back to “Orders” then enter “true” or “false” by vlookup.

	B	W
	Order ID	Returne
	CA-2020-152156	FALSE
2	CA-2020-152156	FALSE
3	CA-2020-138688	FALSE
4	US-2019-108966	FALSE

(columns between these two are hidden)

c. Final combination

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	Row ID	Order ID	Order Date	Ship Date	Ship Mo	Custome	Custome	Segmen	Country	City	State	Postal C	Region	Product ID	Categor	Sub-Cat	Product Name	Sales	Quantity	Discoun	Profit	RegionManager	Returne
1	1	CA-2020-152156	11/8/2020	11/11/2020	Second Cla	CG-12520	Claire Gute	Consumer	United Stat	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96	2	0	41.9136	Fred Suzuki	FALS
1	2	CA-2020-152156	11/8/2020	11/11/2020	Second Cla	CG-12520	Claire Gute	Consumer	United Stat	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	731.94	3	0	219.582	Fred Suzuki	FALS
1	3	CA-2020-138688	6/12/2020	6/16/2020	Second Cla	DV-13045	Darrin Van	Corporate	United Stat	Los Angele	California	90036	West	OFF-LA-10000240	Office Supp	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62	2	0	6.8714	Sadie Pawthorne	FALS
1	4	US-2019-108966	10/11/2019	10/18/2019	Standard C	SO-20335	Sean O'Dou	Consumer	United Stat	Fort Lauder	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.031	Fred Suzuki	FALS

2. Splitting into entities

- Customer(Customer ID, FN, LN, Segment);
- Order(Order ID,Customer ID, Ship Date, Ship Mode, returned, Postal Code);
- Sales(Order ID, Product ID, Quantity, Sales, discount, Profit, cost of sales *which is getting calculated in excel*);
- Product(Product ID, Product Name, Category, Sub-Category)
- Region(Region Code(New created), RegionName *which is actually ‘Region’ in the excel*, ManagerFN, ManagerLN);
- Address(Postal Code, Region Code(New created), State, City, Country *which is actually ‘Country/Region’ in the excel)*

5. Installation: install MySQL Workbench in your computer as per the instruction indicated in this project document

&

6. Database Creation: using your ERD as foundation to create a physical MySQL database “GBC_Superstore”, create tables and fields and add the primary and foreign keys. From MySQL Workbench, generate a printout of the corresponding database schema.

Creating database tables based on ERD and adding primary and foreign keys

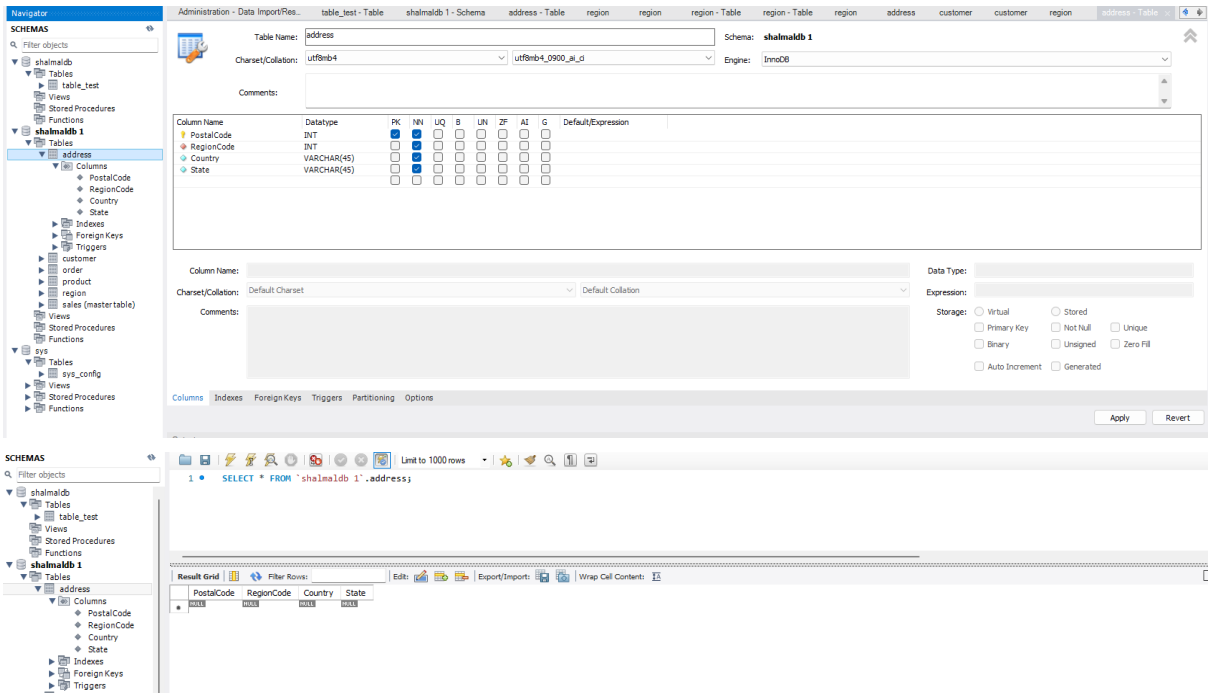


Fig: Address table

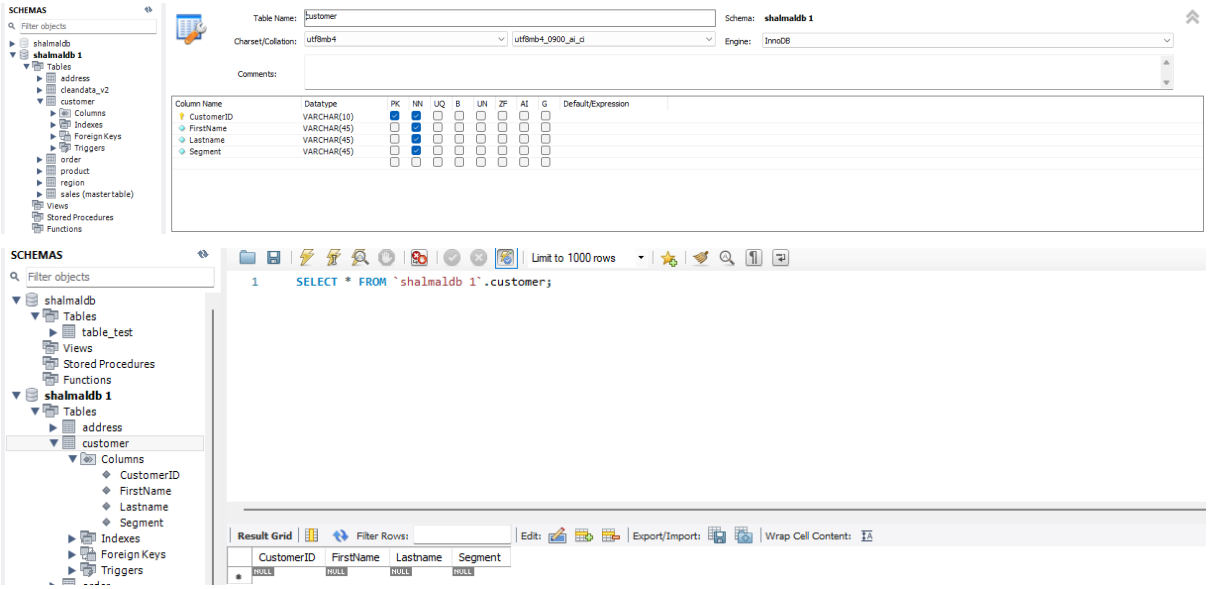


Fig: Customer table

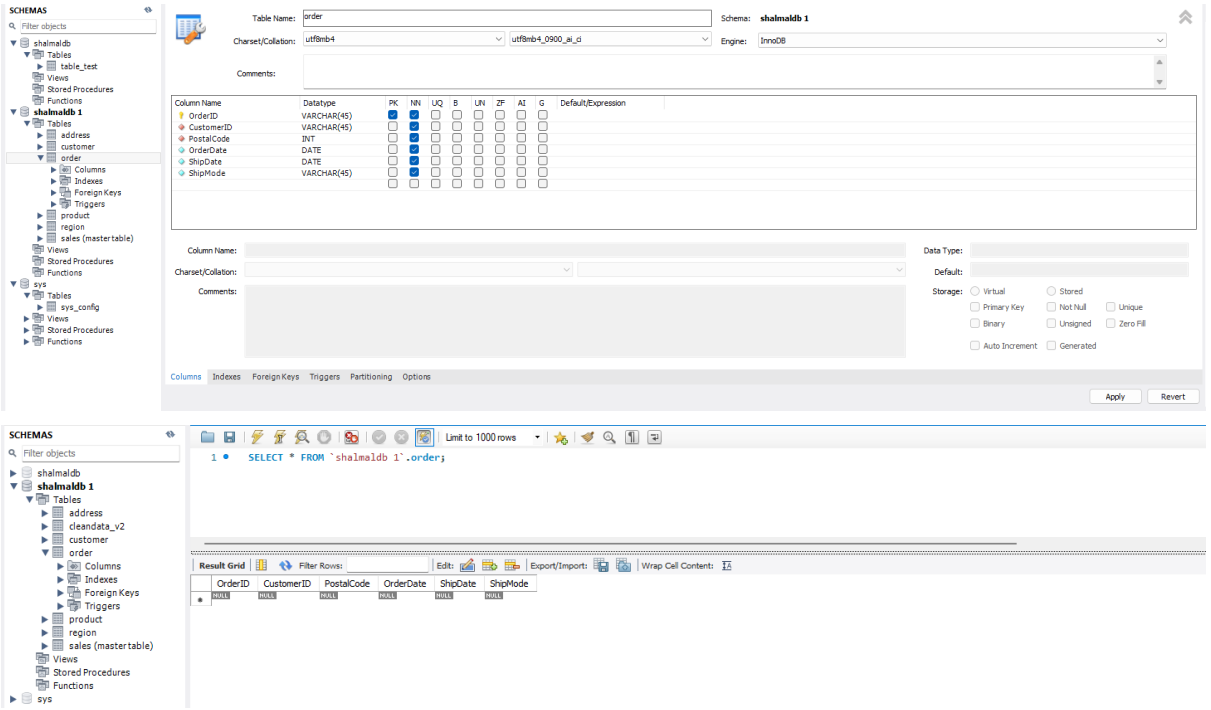


Fig: Order table

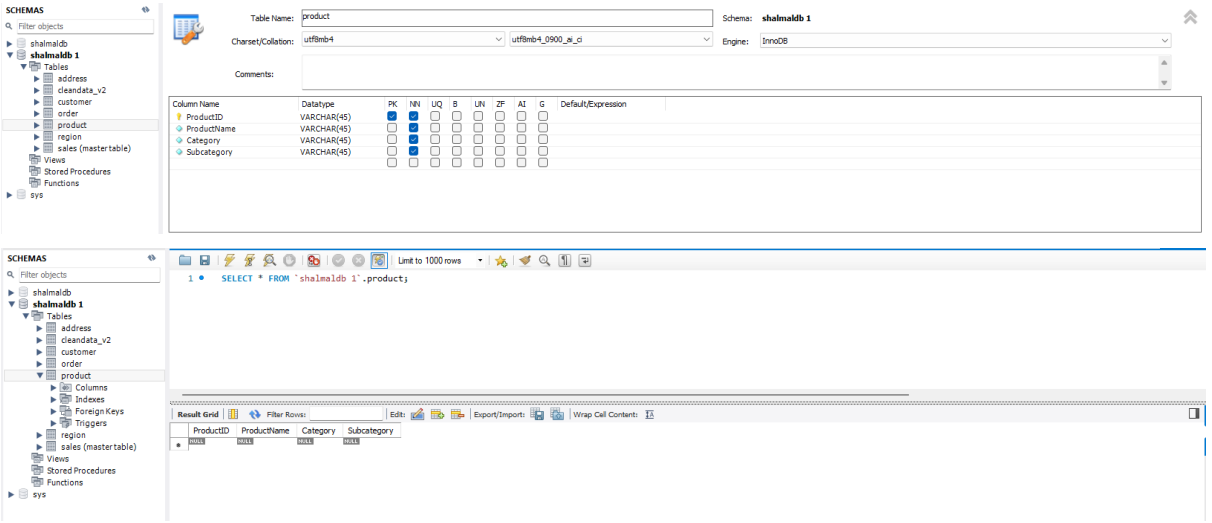


Fig: Product table

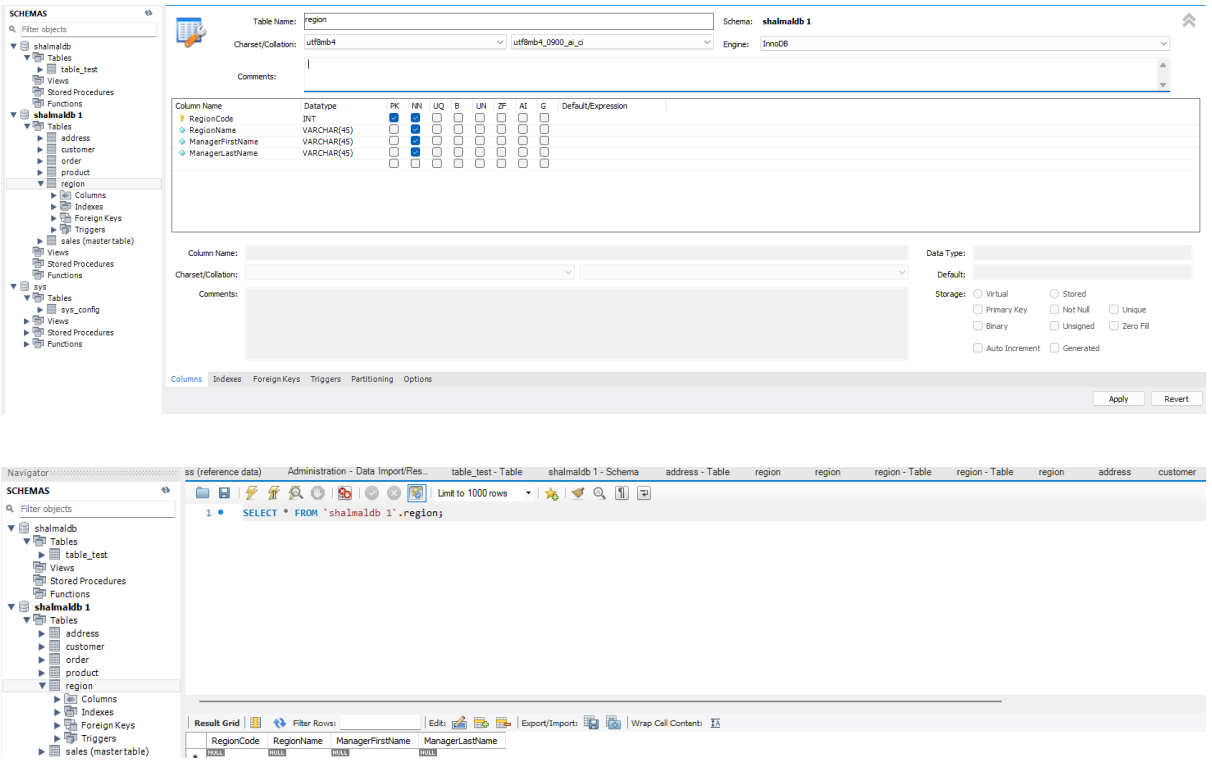


Fig: Region table

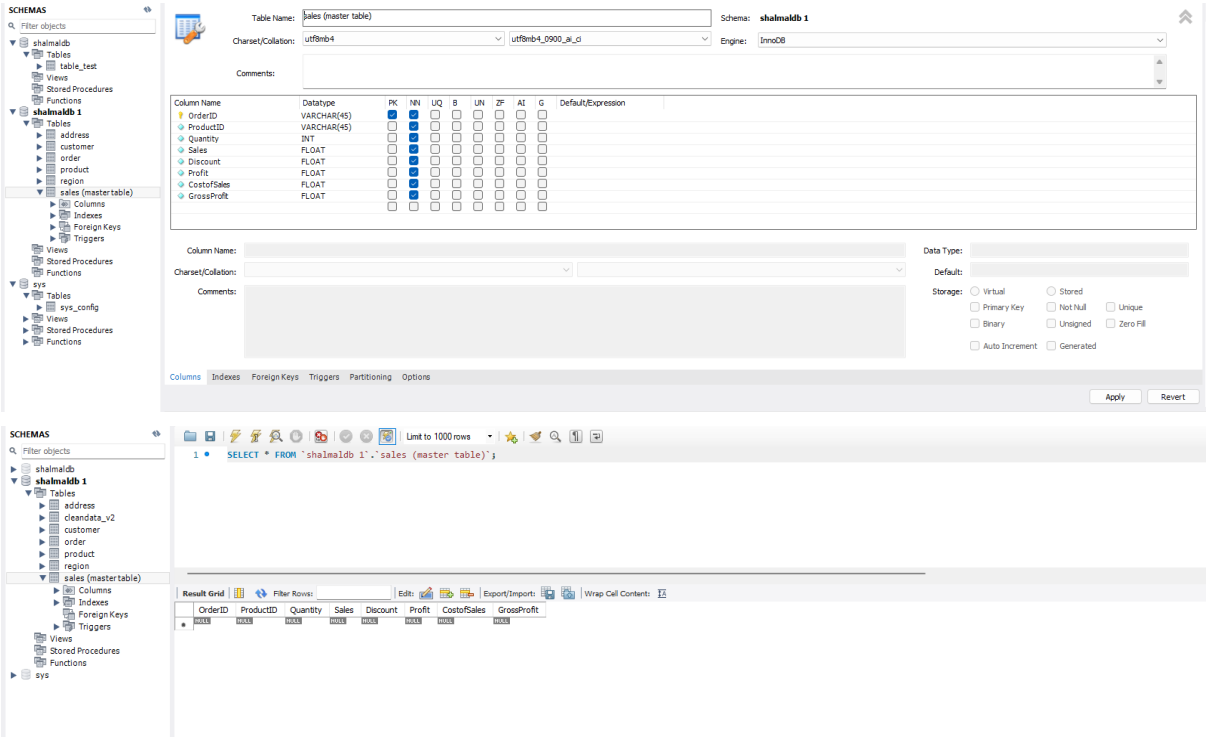
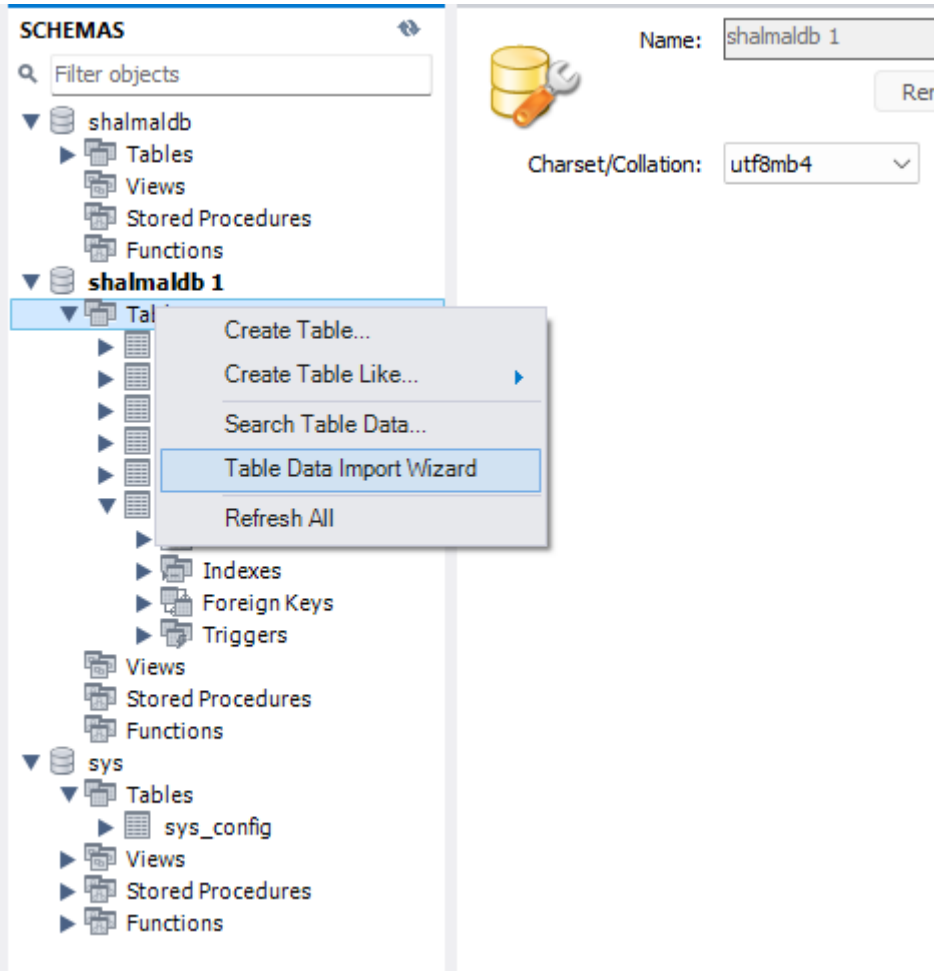


Fig: Sales (Master Table)

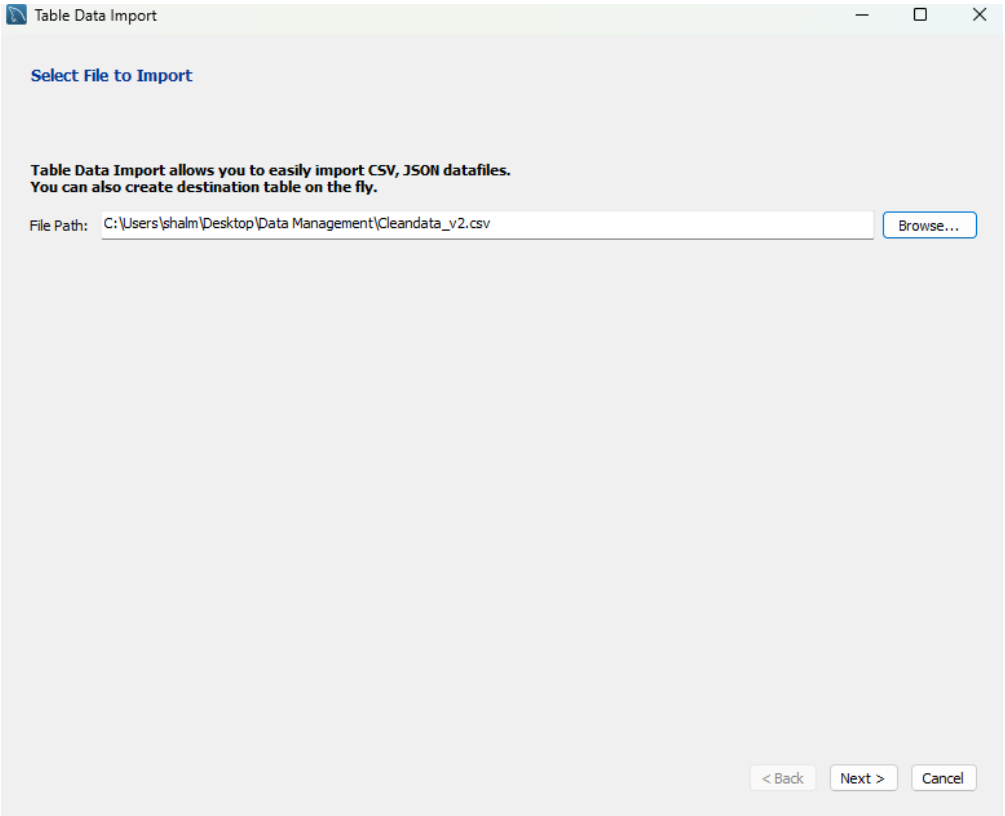
7. ETL: Extract, Transform and Load data from the already cleaned GBC_Superstore.xlsx to MySQL GBC_Superstore tables:

a. generate and export CSV file(s) into MySQL database/tables using SQL scripts and/or
b. use Python to ETL Data from Excel to MySQL tables
c. document the ETL process used for the generation of the MySQL GBC_Superstore database
d. Verify data completeness, data integrity and referential integrity

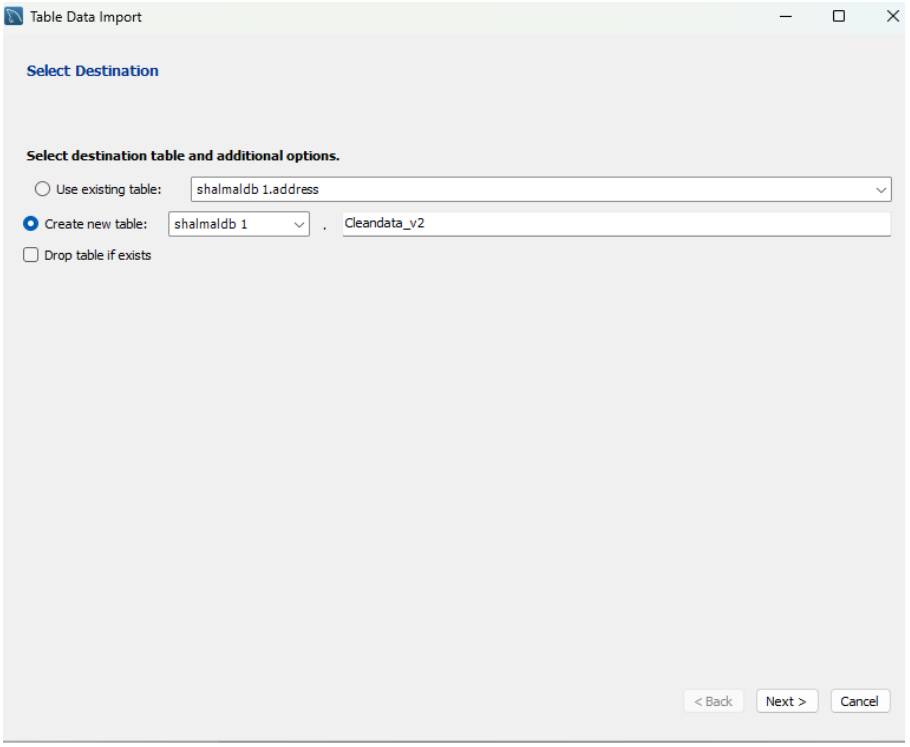
To extract the cleaned data we need to use the Table Data Import Wizard



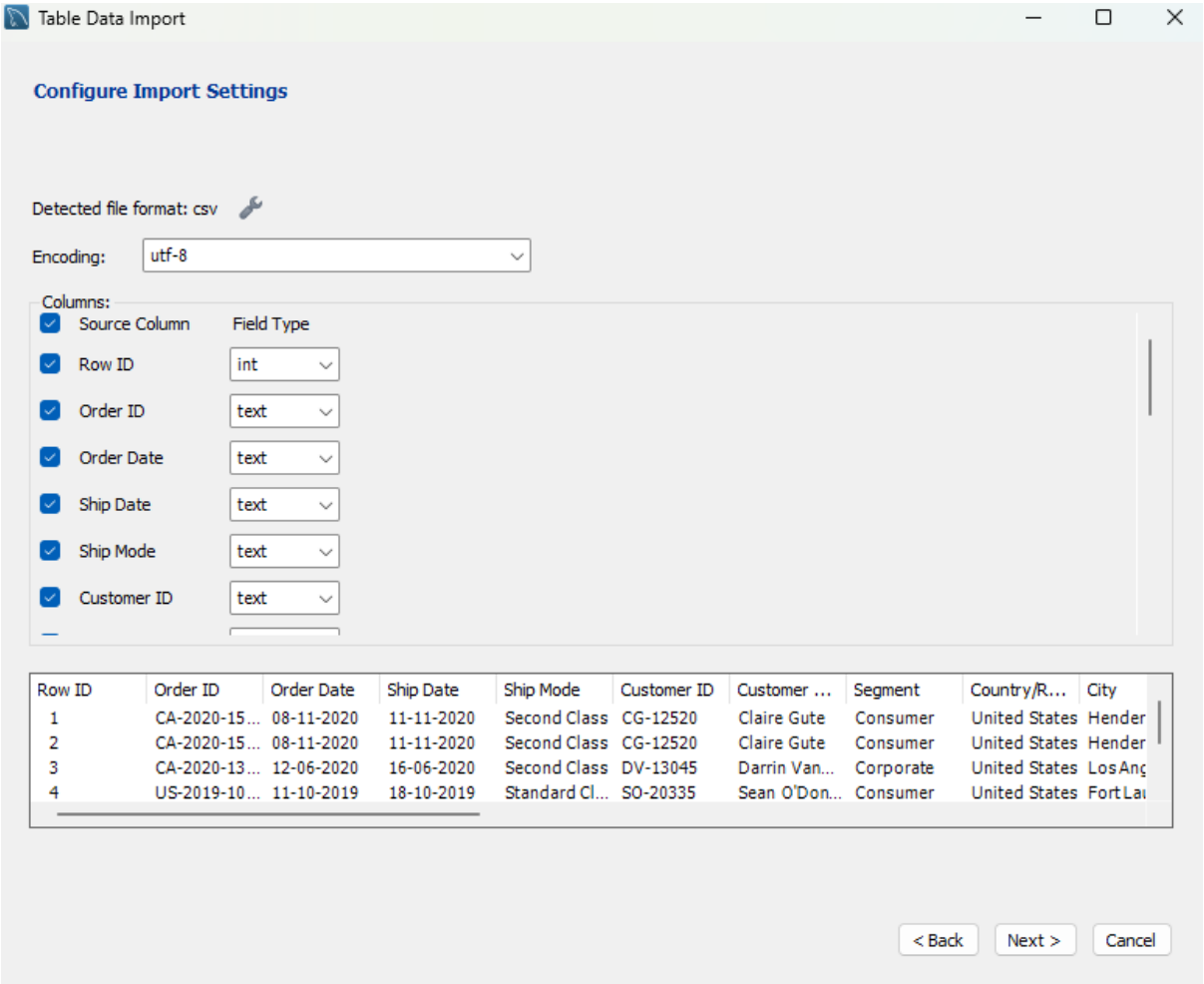
We need to select the path of the .csv file



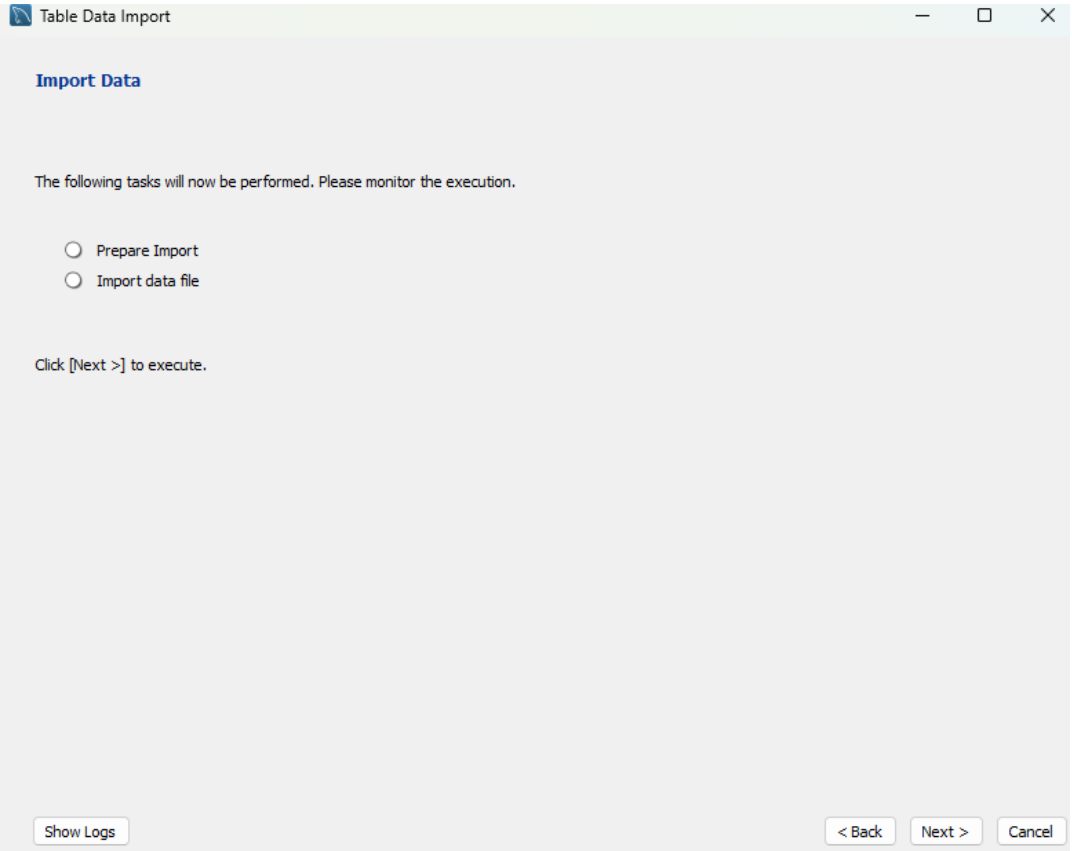
We need to select whether we want a new table for the dataset or we want to import it to a specific table



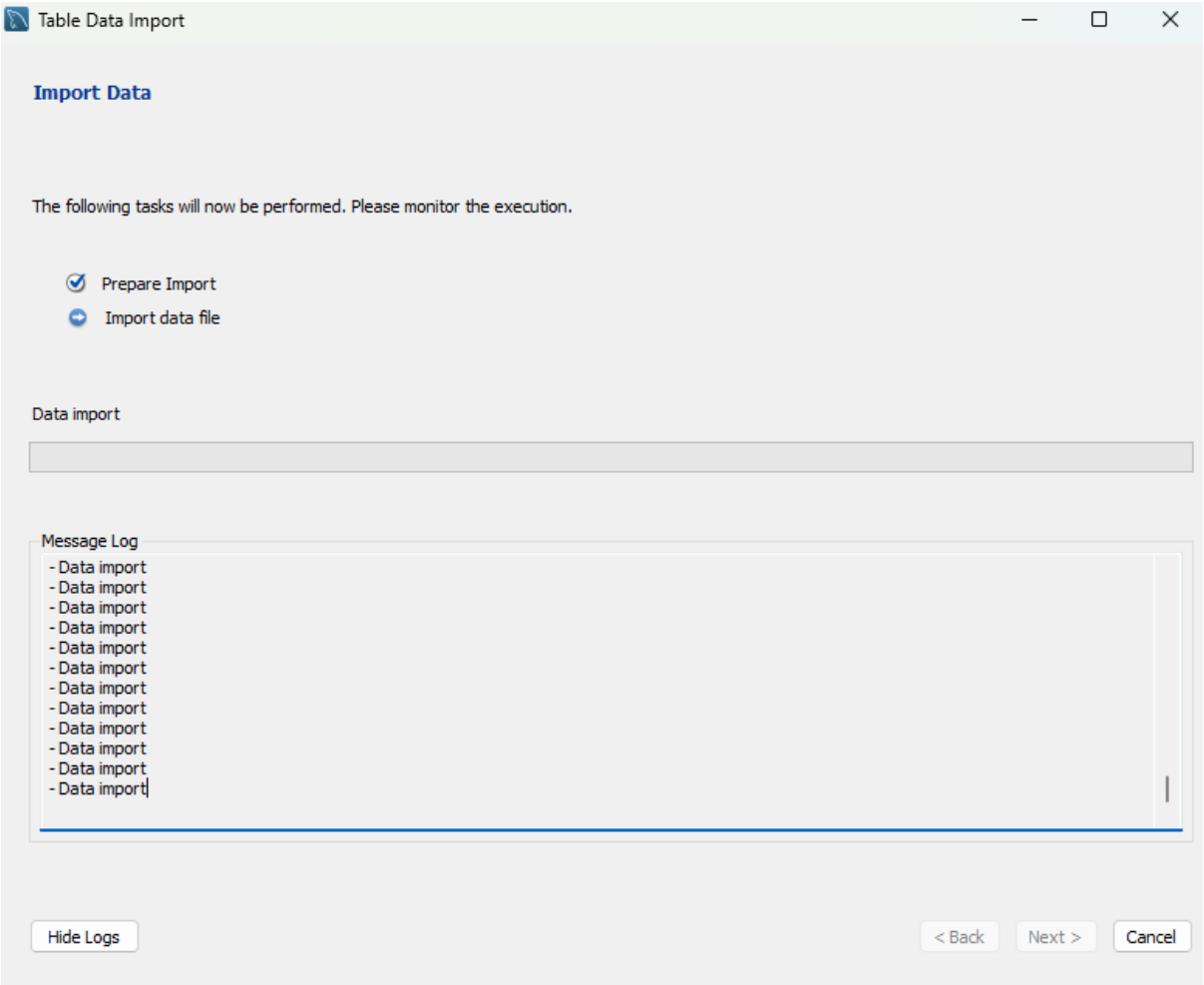
We have to select the encoding format and the datatypes of all the columns of the cleaned dataset.



By pressing next the wizard will prepare the data for import



The data is being imported to SQL table



Printout of database Schema:

It matches with our *Logical-level ERD.*