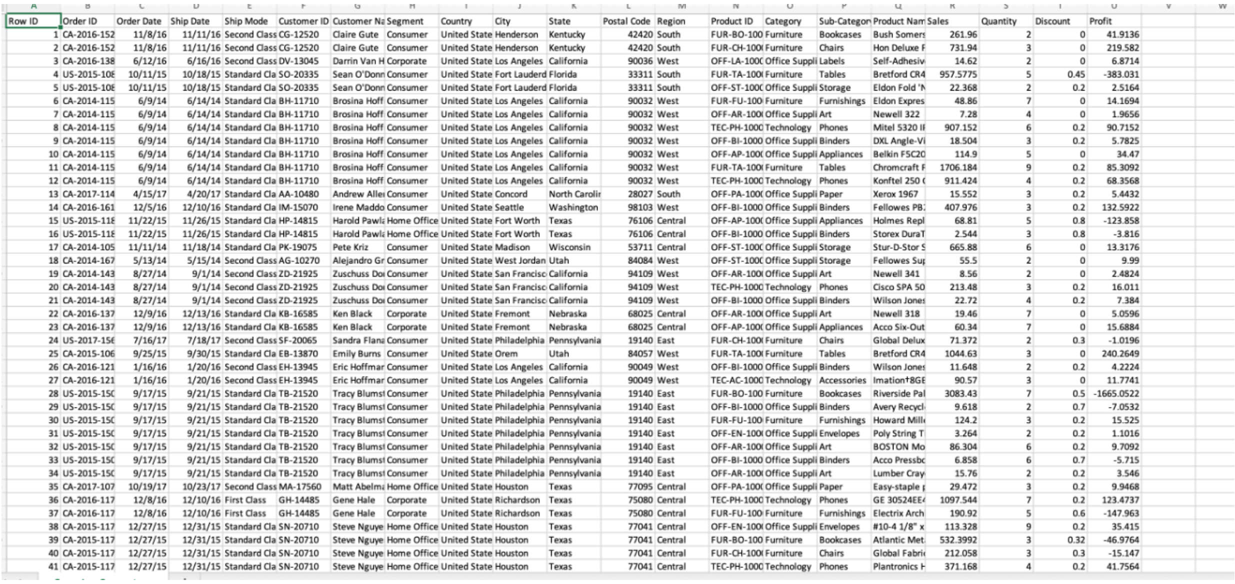
Data Modeling | Dimensional Modeling | ER Diagram | Star Schema | Snowflake Schema

- For this we will use a sample Superstore Data set that has the following information:
  - Metadata:
    - Row ID => Unique ID for each row.
    - Order ID => Unique Order ID for each Customer.
    - Order Date => Order Date of the product.
    - Ship Date => Shipping Date of the Product.
    - Ship Mode=> Shipping Mode specified by the Customer.
    - Customer ID => Unique ID to identify each Customer.
    - Customer Name => Name of the Customer.
    - Segment => The segment where the Customer belongs.
    - Country => Country of residence of the Customer.
    - City => City of residence of of the Customer.
    - State => State of residence of the Customer.
    - Postal Code => Postal Code of every Customer.
    - Region => Region where the Customer belong.
    - Product ID => Unique ID of the Product.
    - Category => Category of the product ordered.
    - Sub-Category => Sub-Category of the product ordered.
    - Product Name => Name of the Product
    - Sales => Sales of the Product.
    - Quantity => Quantity of the Product.
    - Discount => Discount provided.
    - Profit => Profit/Loss incurred.
  - Kashif - Sample - Superstore.csv.zip
- This data set is about a Superstore giant who is trying to understand what works best for them.
- They would like to understand which products, regions, categories and customer segments they should target or avoid.



Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Na	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Categor	Product Nam	Sales	Quantity	Discount	Profit
1	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-BO-100	Furniture	Bookcases	Bush Somers	261.96	2	0	41.9136
2	CA-2016-152	11/8/16	11/11/16	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-CH-1000	Furniture	Chairs	Hon Deluxe F	731.94	3	0	219.582
3	CA-2016-138	6/12/16	6/16/16	Second Class	DV-13045	Darrin Van H	Corporate	United State	Los Angeles	California	90036	West	OFF-LA-1000	Office Suppli	Labels	Self-Adhesiv	14.62	2	0	6.8714
4	US-2015-108	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	FUR-TA-1000	Furniture	Tables	Bretford CR4	957.5775	5	0.45	-383.031
5	US-2015-108	10/11/15	10/18/15	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	OFF-ST-1000	Office Suppli	Storage	Eldon Fold 'N	22.368	2	0.2	2.5164
6	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-FU-1000	Furniture	Furnishings	Eldon Expres	48.86	7	0	14.1694
7	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-AR-1000	Office Suppli	Art	Newell 322	7.28	4	0	1.9656
8	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Mitel 5320 IF	907.152	6	0.2	90.7152
9	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-BI-1000	Office Suppli	Binders	DXL Angle-Vi	18.504	3	0.2	5.7825
10	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	OFF-AP-1000	Office Suppli	Appliances	Belkin F5C20	114.9	5	0	34.47
11	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	FUR-TA-1000	Furniture	Tables	Chromcraft F	1706.184	9	0.2	85.3092
12	CA-2014-115	6/9/14	6/14/14	Standard Cla	BH-11710	Brosina Hoff	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-1000	Technology	Phones	Konftel 250 C	911.424	4	0.2	68.3568
13	CA-2017-114	4/15/17	4/20/17	Standard Cla	AA-10480	Andrew Aller	Consumer	United State	Concord	North Carolir	28027	South	OFF-PA-1000	Office Suppli	Paper	Xerox 1967	15.552	3	0.2	5.4432
14	CA-2016-161	12/5/16	12/10/16	Standard Cla	IM-15070	Irene Maddo	Consumer	United State	Seattle	Washington	98103	West	OFF-BI-1000	Office Suppli	Binders	Fellowes PB	407.976	3	0.2	132.5922
15	US-2015-118	11/22/15	11/26/15	Standard Cla	HP-14815	Harold Pawla	Home Office	United State	Fort Worth	Texas	76106	Central	OFF-AP-1000	Office Suppli	Appliances	Holmes Repl	68.81	5	0.8	-123.858
16	US-2015-118	11/22/15	11/26/15	Standard Cla	HP-14815	Harold Pawla	Home Office	United State	Fort Worth	Texas	76106	Central	OFF-BI-1000	Office Suppli	Binders	Storex DuraT	2.544	3	0.8	-3.816
17	CA-2014-105	11/11/14	11/18/14	Standard Cla	PK-19075	Pete Kriz	Consumer	United State	Madison	Wisconsin	53711	Central	OFF-ST-1000	Office Suppli	Storage	Stur-D-Stor S	665.88	6	0	13.3176
18	CA-2014-167	5/13/14	5/15/14	Second Class	AG-10270	Alejandro Gr	Consumer	United State	West Jordan	Utah	84084	West	OFF-ST-1000	Office Suppli	Storage	Fellowes Sup	55.5	2	0	9.99
19	CA-2014-143	8/27/14	9/1/14	Second Class	ZD-21925	Zuschuss Do	Consumer	United State	San Francisc	California	94109	West	OFF-AR-1000	Office Suppli	Art	Newell 341	8.56	2	0	2.4824
20	CA-2014-143	8/27/14	9/1/14	Second Class	ZD-21925	Zuschuss Do	Consumer	United State	San Francisc	California	94109	West	TEC-PH-1000	Technology	Phones	Cisco SPA 50	213.48	3	0.2	16.011
21	CA-2014-143	8/27/14	9/1/14	Second Class	ZD-21925	Zuschuss Do	Consumer	United State	San Francisc	California	94109	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	22.72	4	0.2	7.384
22	CA-2016-137	12/9/16	12/13/16	Standard Cla	KB-16585	Ken Black	Corporate	United State	Fremont	Nebraska	68025	Central	OFF-AR-1000	Office Suppli	Art	Newell 318	19.46	7	0	5.0596
23	CA-2016-137	12/9/16	12/13/16	Standard Cla	KB-16585	Ken Black	Corporate	United State	Fremont	Nebraska	68025	Central	OFF-AP-1000	Office Suppli	Appliances	Acco Six-Out	60.34	7	0	15.6884
24	US-2017-156	7/16/17	7/18/17	Second Class	SF-20065	Sandra Flana	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-CH-1000	Furniture	Chairs	Global Delux	71.372	2	0.3	-1.0196
25	CA-2015-106	9/25/15	9/30/15	Standard Cla	EB-13870	Emily Burns	Consumer	United State	Orem	Utah	84057	West	FUR-TA-1000	Furniture	Tables	Bretford CR4	1044.63	3	0	240.2649
26	CA-2016-121	1/16/16	1/20/16	Second Class	EH-13945	Eric Hoffmar	Consumer	United State	Los Angeles	California	90049	West	OFF-BI-1000	Office Suppli	Binders	Wilson Jones	11.648	2	0.2	4.2224
27	CA-2016-121	1/16/16	1/20/16	Second Class	EH-13945	Eric Hoffmar	Consumer	United State	Los Angeles	California	90049	West	TEC-AC-1000	Technology	Accessories	Imation†8GE	90.57	3	0	11.7741
28	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-BO-100	Furniture	Bookcases	Riverside Pal	3083.43	7	0.5	-1665.0522
29	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-BI-1000	Office Suppli	Binders	Avery Recycl	9.618	2	0.7	-7.0532
30	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	FUR-FU-1000	Furniture	Furnishings	Howard Mille	124.2	3	0.2	15.525
31	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-EN-1000	Office Suppli	Envelopes	Poly String T	3.264	2	0.2	1.1016
32	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-AR-1000	Office Suppli	Art	BOSTON Mo	86.304	6	0.2	9.7092
33	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-BI-1000	Office Suppli	Binders	Acco Pressbo	6.858	6	0.7	-5.715
34	US-2015-150	9/17/15	9/21/15	Standard Cla	TB-21520	Tracy Blumst	Consumer	United State	Philadelphia	Pennsylvania	19140	East	OFF-AR-1000	Office Suppli	Art	Lumber Cray	15.76	2	0.2	3.546
35	CA-2017-107	10/19/17	10/23/17	Second Class	MA-17560	Matt Abelma	Home Office	United State	Houston	Texas	77095	Central	OFF-PA-1000	Office Suppli	Paper	Easy-staple p	29.472	3	0.2	9.9468
36	CA-2016-117	12/8/16	12/10/16	First Class	GH-14485	Gene Hale	Corporate	United State	Richardson	Texas	75080	Central	TEC-PH-1000	Technology	Phones	GE 30524EE4	1097.544	7	0.2	123.4737
37	CA-2016-117	12/8/16	12/10/16	First Class	GH-14485	Gene Hale	Corporate	United State	Richardson	Texas	75080	Central	FUR-FU-1000	Furniture	Furnishings	Electrix Arch	190.92	5	0.6	-147.963
38	CA-2015-117	12/27/15	12/31/15	Standard Cla	SN-20710	Steve Nguye	Home Office	United State	Houston	Texas	77041	Central	OFF-EN-1000	Office Suppli	Envelopes	#10-4 1/8" x	113.328	9	0.2	35.415
39	CA-2015-117	12/27/15	12/31/15	Standard Cla	SN-20710	Steve Nguye	Home Office	United State	Houston	Texas	77041	Central	FUR-BO-100	Furniture	Bookcases	Atlantic Met	532.3992	3	0.32	-46.9764
40	CA-2015-117	12/27/15	12/31/15	Standard Cla	SN-20710	Steve Nguye	Home Office	United State	Houston	Texas	77041	Central	FUR-CH-1000	Furniture	Chairs	Global Fabri	212.058	3	0.3	-15.147
41	CA-2015-117	12/27/15	12/31/15	Standard Cla	SN-20710	Steve Nguye	Home Office	United State	Houston	Texas	77041	Central	TEC-PH-1000	Technology	Phones	Plantronics H	371.168	4	0.2	41.7564

- This can be interpreted as a **De-normalized Database** where everything is stored in One Big Table.
- Leading to **Data Duplication Issue**, which increases cost and reduces performance.
- For designing a Normalized Data Warehouse system with the above data set, we will use Dimensional Modeling.
    - **Dimensional Modeling**: Is nothing but organizing the data in a structured way by creating **Fact** (measurements) and **Dimension** (contextual information) tables.

    - **Benefits of Dimensional Modeling:**
        - Improved Query Performance
        - Simplified Data Analysis
        - Enhanced Data Visualization

    - **Fact Tables**:
        - Consists of quantifiable aspects of data. From the above sample data set, a fact table will consist of sales amount, product quantities, discount, profit.
        - Represents actual values or metrics that are being analyzed.
        - Store facts like actual business data (Profit, Sales, Revenue), are linked to Dimension tables.

    - **Dimension Tables**:
        - Stores contextual information.
        - Are descriptive in nature. Like a Lookup Table. Categorical Data.
        - Provides context to the measurements.
        - From the above sample data set, information related to the Product, Customer and Location will be stored in respective Dimension Tables.

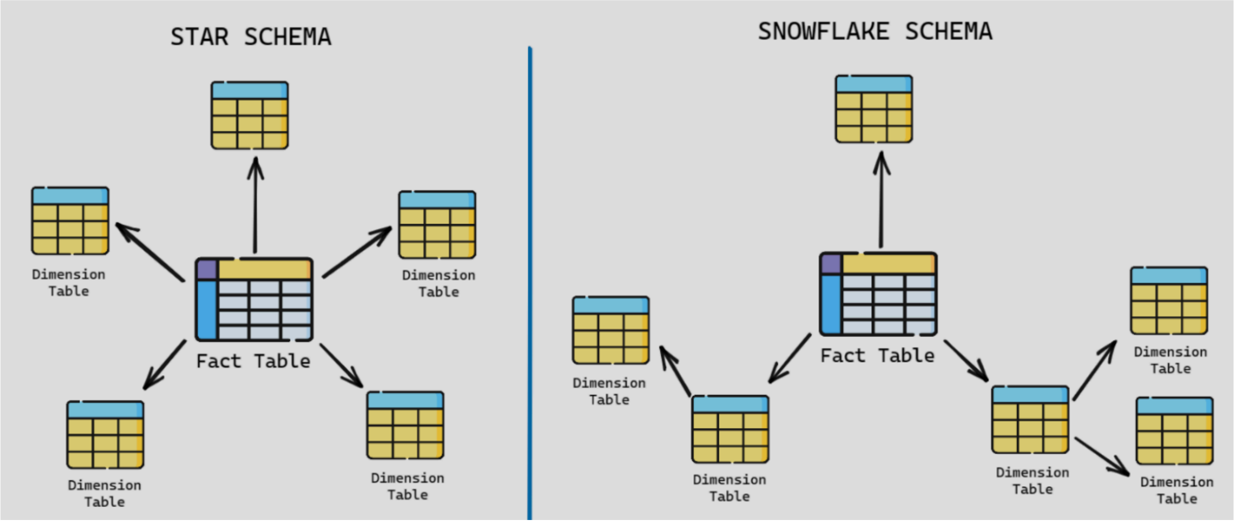
Dimensional Modeling

- Structural approaches to building a Dimension Model:
  - Star Schema
  - Snowflake Schema

- **Star Schema**:
  - Dimension Tables are just 1 level away from Fact Table
  - Fewer Database joins
  - More Database storage
  - Hierarchies are in one dimension table
  - Used for simple DB Design
  - Higher Query Performance

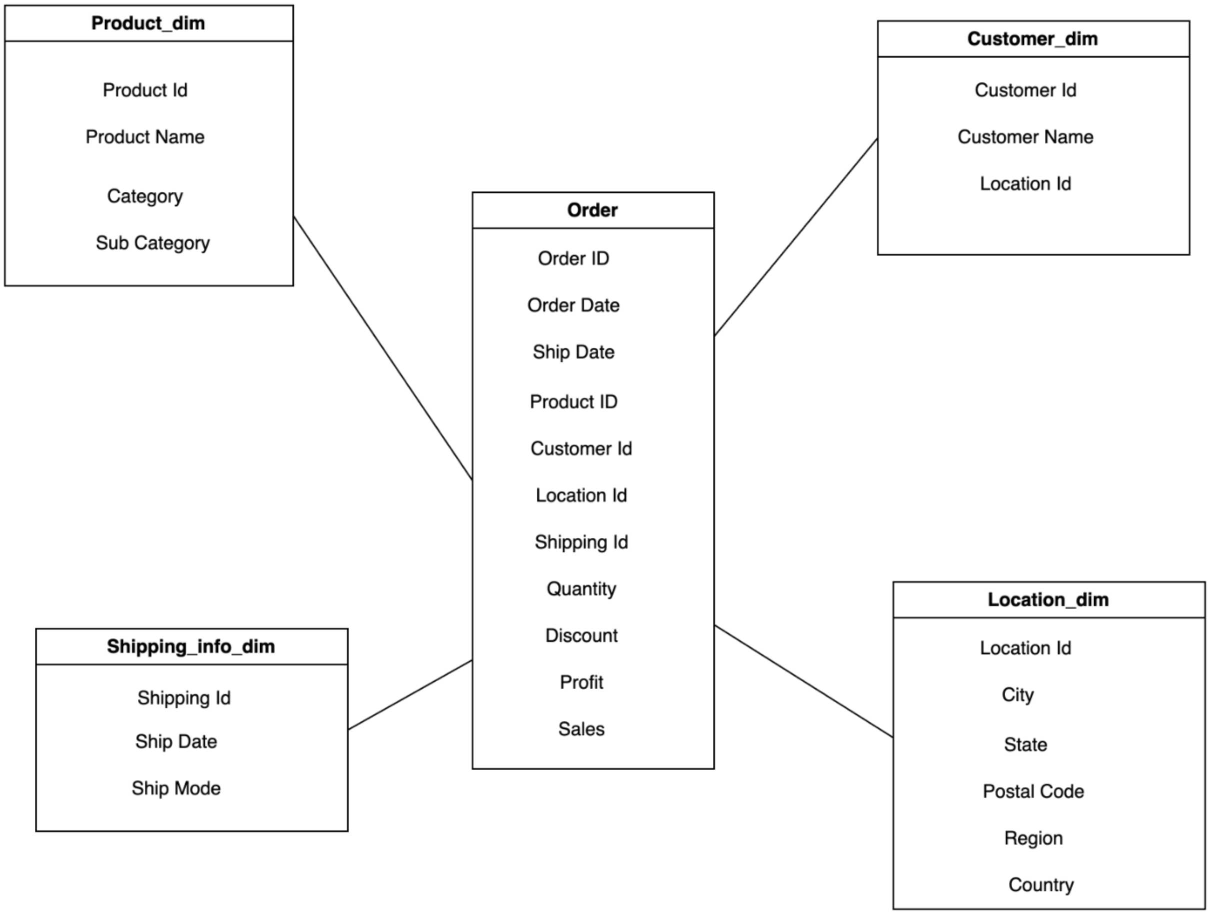
- **Snowflake Schema**:
  - Dimension Tables are 1 or more levels away from Fact Table.
  - Hierarchies are divided into multiple Dimension Tables.
  - More Database Joins
  - Less Database storage
  - Used for Very Complex DB Design

- **Choosing a right schema depends on**:
  - Storage Requirements
  - Complexity of the Database Relationships
  - No of Joins Needed for Data Analysis

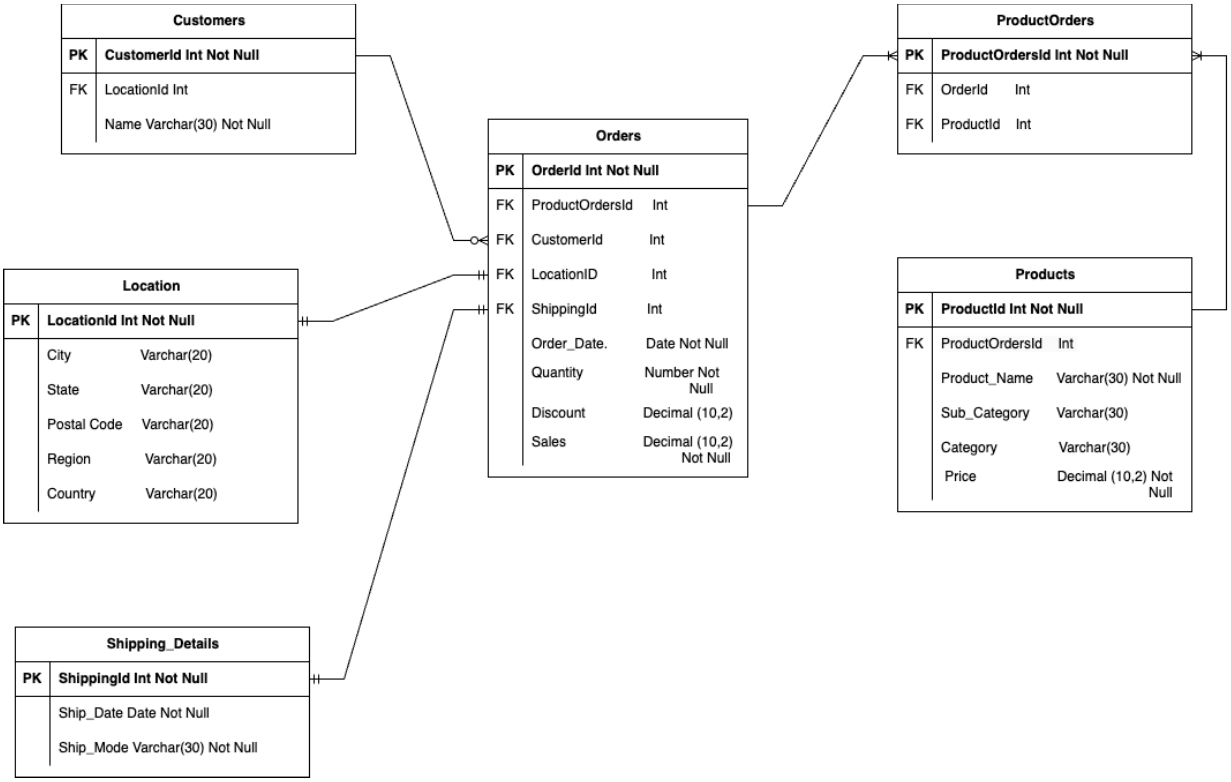


Data Modeling Lifecycle:

1. In the first phase called as **Conceptual Modeling**, lets create Fact and Dimension Tables and link them together in a **Star Schema**.
    a. This phase is more like getting a look inside view of the data and keeping the data as near to the real world as possible to get more insights about the data.
    b. We opted for Star Schema so that we only have dimension tables 1 level away from fact table and that there are fewer database joins.



2. In the second phase of Dimension Modeling called as **Logical Modeling**, lets create an ER Diagram.
    a. Database to be used is Redshift for this sample use case.
    b. **ER Modeling**:



Customers

PK **CustomerId Int Not Null**

FK LocationId Int

Name Varchar(30) Not Null

Orders

PK **OrderId Int Not Null**

FK ProductOrdersId Int

FK CustomerId Int

FK LocationID Int

FK ShippingId Int

Order_Date. Date Not Null

Quantity Number Not Null

Discount Decimal (10,2)

Sales Decimal (10,2) Not Null

ProductOrders

PK **ProductOrdersId Int Not Null**

FK OrderId Int

FK ProductId Int

Location

PK **LocationId Int Not Null**

City Varchar(20)

State Varchar(20)

Postal Code Varchar(20)

Region Varchar(20)

Country Varchar(20)

Products

PK **ProductId Int Not Null**

FK ProductOrdersId Int

Product_Name Varchar(30) Not Null

Sub_Category Varchar(30)

Category Varchar(30)

Price Decimal (10,2) Not Null

Shipping_Details

PK **ShippingId Int Not Null**

Ship_Date Date Not Null

Ship_Mode Varchar(30) Not Null

3. The third phase would be creating Database scripts to create tables and load the data in it.
    a. You could use Tools to generate the script from your ER Diagram (Lucid-chart for example) or create them manually.