ORION STAR “SWORD TEAM”

Mahendra Bairi
31031490

PRESENTATION CONTENT

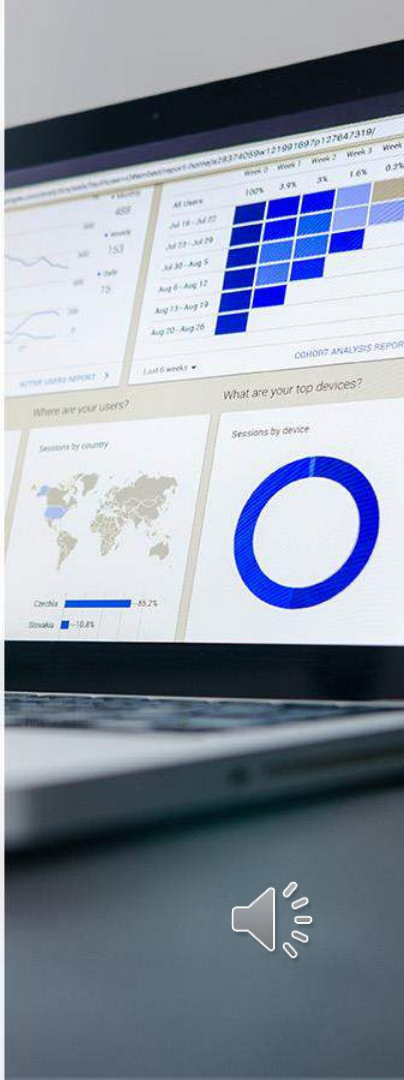
- Introduction
- Business Requirements
- Problem Statement
- Justification for Business Question
- Development Tools
- Methodology
- Datasets
- Justification for Chosen Datasets
- Data Cleaning
- Data Loading
- Data visualization
- Requirements
- References

INTRODUCTION

- Orion Star is an international retail company which is selling sports and outdoor products, headquarter is located at United States and also has multiple retail stores which are located in several other countries.
- Orion Star sells it’s products through physical retail stores, mail-order catalogues, and the internet.
- It has Orion Star club which provides special offers and discounts for the enrolled customers.
- The sales data provided between the years 1998 to 2002 by Orion Star club members purchases.

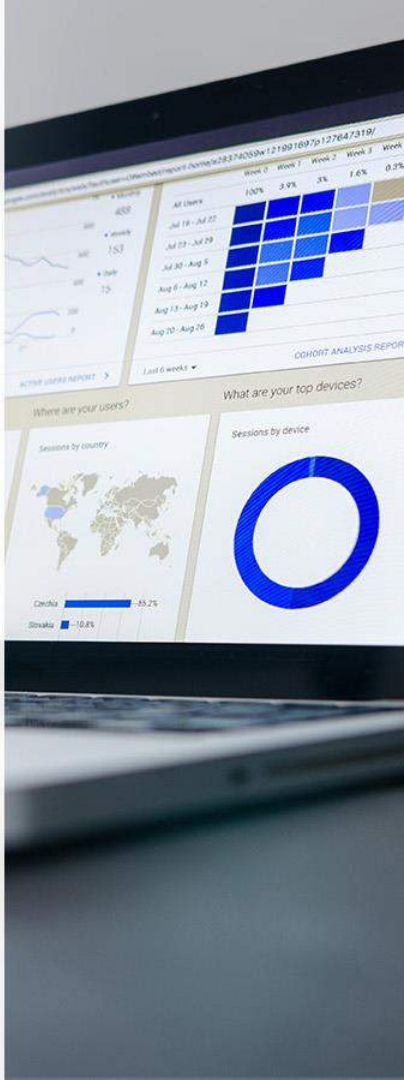
BUSINESS REQUIREMENTS

- The sponsor has noticed many vulnerabilities with business intelligence which must be solved.
- The primary goal is to answer the sponsor’s needs and solve the difficulties that have been stated, In order to do that we need to design and build an acceptable data structure which can generate standard reports.
- And also user needs and interface, data warehouse architecture, identification, using suitable tools and methodology and business strategy.



PROBLEM STATEMENT ???

Overall the project is to understand the requirements of the sponsor and provide a solution that meets their needs. Mainly the project should focus on designing and building a suitable data structure that allows for the efficient and effective storage and analysis of data. Also the company has provided sales data and few business questions which can be solved by using Big Data Techniques to get insights for their business operation.



JUSTIFICATION FOR BUSINESS QUESTIONS

1. **Who were the last month's top ten customers by sales value?**

   This brings up the most valuable customers with what products they purchased and kind of services they utilized which dispense valuable insights into customer base, those will help to increase revenue by developing strategies, customer satisfaction in order to achieve business growth.

2. **What was the most profitable product of 2002?**

   By knowing the most profitable product for particular year can bring insights for future sales of business, accordingly plan inventory and resources.

3. **What was the total number of products purchased in 2002 by customer type?**

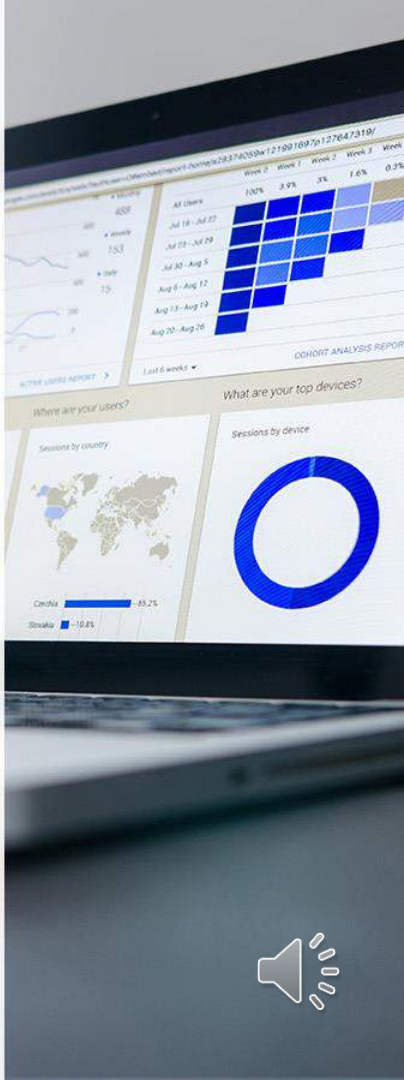
   Understanding the purchasing patterns of different types of customers enables you to offer better rates, better customer care, and more tailored products for certain types of customers, which will help your business expand.

4. **Which customer type has generated the least amount of income in 2002?**

   By identifying the customer type that obtained the least income, investments will be prioritized to raise revenue by retaining consumers from that group.

5. **What were the total sales over time by country by product group?**

   From above question it will provide valuable insights like most selling products in particular country from that business and increase their production and marketing in that particular region for those products.



DEVELOPMENT TOOLS

- Rstudio : R is open source and it is an integrated development environment(IDE). It is a dedicated tool for statistical programming language and data science with great communication, where python is used for multipurpose such as coding, analysis. R is integrated with some other tools for data analysis and also includes lot packages with many functions used for data validation, extraction, quality checks, transformation, and cleaning. These are the advantages r have while comparing with other tools
- Hive : Comparing with other tools, It is built on top of apache Hadoop and it is integrated with Hadoop and designed to work with large-datasets easily like petabytes of data. And also it has SQL interface, In addition it reduces the complexity of mapreduce framework and also used to read, write and manage the data for analysis.
- Tableau: It is a data visualization tools which provide drag and drop interface to create graphs and tables of data that makes analysis faster, easier and industry standard because it handle large volumes of data where other tools are not.

Figure: Used tools for project development

- In this project I have used the mentioned tools to develop the project and get insights for business questions given by organization.

METHODOLOGY

- To built and design a data warehouse here I choose Kimball's bottom-up approach.
- In 1990's the data warehouse expert ralph Kimball developed this Kimball approach.
- As per business requirement this model go after bottom-up approach to design data warehouse architecture with data marts first, compared with Top-down approach is complex and designed to be enterprise wide.
- This method allows data from multiple data sources and follows ETL process in order to create data warehouse staging area. In this architecture data warehouse designed with data marts divided into, facts which is having numerical transactional data such as quantity, price, and dimensions with categories or reference information which supports facts such as product, time.

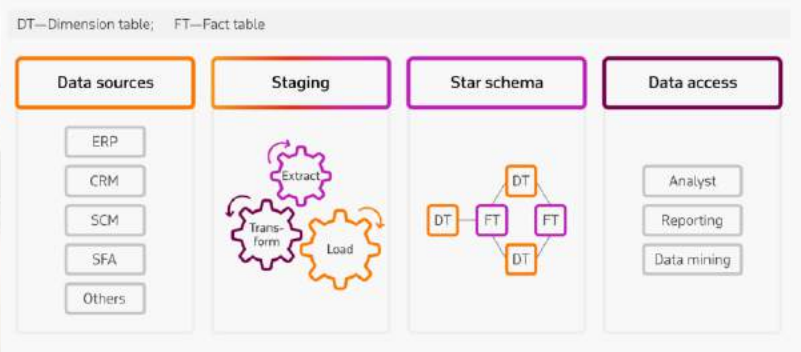
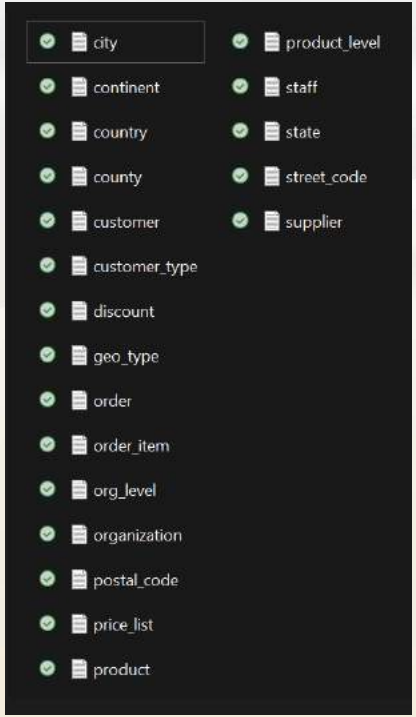


Figure: Kimball's approach to design a data warehouse

Justification for choosing bottom-up method

- Initial setup is easy to create data warehouse, no normalization required.
- Querying and analysis are easy with data dimensional model
- Easy to understand and make informed decision making for business users.
- Multiple star schemas can be designed to produce business requirement insights
- Early stage time and cost is less.
- ETL transforms data before processing into data warehouse, while ELT transforms raw data to DWH which expose private data and not much standardized.

DATASETS



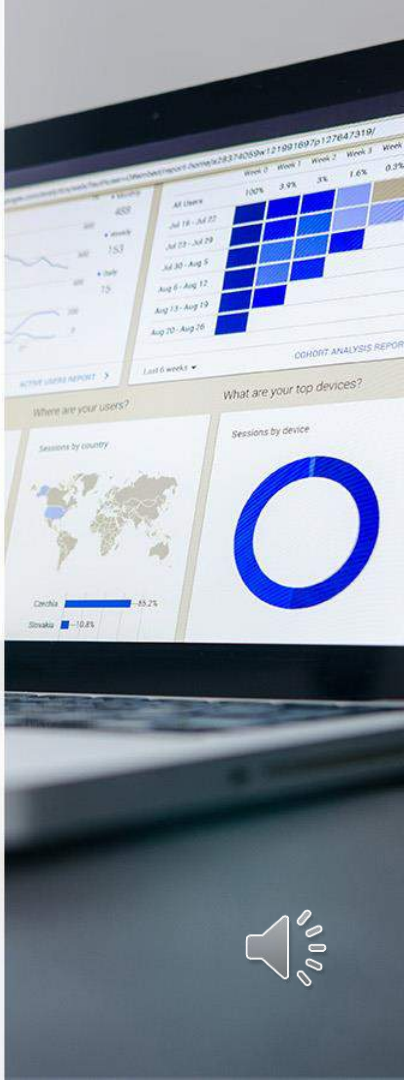

```
#importing the txt file

customer <- read.csv("C:/User
customer_type <- read.csv("C:
order <- read.csv("C:/Users/M
order_item <- read.csv("C:/Us
country <- read.csv("C:/Users
price_list <- read.csv("C:/Us
product <- read.csv("C:/Users
product_level <- read.csv("C:

#view the data sets
```

- Data source is from organization provided by IS department. The provided data must be transform cleanse and must load. Out of 20 datasets after analyzing the datasets I have chosen only 8 datasets in order to get reports for provided business questions. In following slide justification is provided.

JUSTIFICATION FOR CHOSEN DATASETS

To answer the business questions I have mentioned the attributes below that to particular datasets.

1. **Who were the last month's top ten customers by sales value?**
   - customers name from customer table
   - order date from order table
   - total sales price from order_item table
2. **What was the most profitable product of 2002?**
   - Product_Name from Product table
   - Order_date(year) from order
   - Total_retail_price, quantity from order_item
   - Unit cost price from price_list
3. **What was the total number of products purchased in 2002 by customer type?**
   - Customer_type from customer_type table
   - Quantity from order_item table
   - Order_date(year) from order
4. **Which customer type has generated the least amount of income in 2002?**
   - Customer_type from customer_type table
   - Order_date(year) from order
   - Total_retail_price from order_item
5. **What were the total sales over time by country by product group?**
   - Country from country table
   - Product group from product table
   - Order_date from order
   - Total_retail_price from order_item



DATA CLEANING

- By using R-studio I extracted and transformed dirty data by using multiple packages to utilize those functions for cleaning.

Mainly the operations done on data are

- Class conversion
- Missing data
- Duplicates
- Rearranging the columns in order
- Merging the datasets
- Special characters/symbols
- Crazy character such as â í ü Â á ą ę ś ć

DATA EXTRACTION

- By using R-studio data extraction has done

```
customer <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarData/
customer_type <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStar
order <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarData/ord
order_item <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarDat
country <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarData/c
price_list <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarDat
product <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStarData/p
product_level <- read.csv("C:/Users/Mahendra/OneDrive/Desktop/Re assesment  ADMP/OrionStar
```

Customer_Type_ID	Customer_Type	Customer_Group_ID	Customer_Group
1010	Orion Club members inactive	10	Orion Club members
1020	Orion Club members low activity	10	Orion Club members
1030	Orion Club members medium activity	10	Orion Club members
1040	Orion Club members high activity	10	Orion Club members
2010	Orion Club Gold members low activity	20	Orion Club Gold members
2020	Orion Club Gold members medium activity	20	Orion Club Gold members
2030	Orion Club Gold members high activity	20	Orion Club Gold members
3010	Internet/Catalog Customers	30	Internet/Catalog Customers

Customer_ID	Country	Gender	Personal_ID	Customer_Name	Customer_FirstName	Customer_LastName	Birth_Date	Customer_Address	Street_
1	FR	M	NA	Albert Collet	Albert	Collet	24NOV1940	Square Edouard VII 1	35001
2	ES	F	NA	Mercedes Martínez	Mercedes	Martínez	15JAN1955	Edificio 2	83001
3	IT	M	NA	Pier Egidio Boeris	Pier Egidio	Boeris	01JUL1970	Via M. Di Montesole 3	48001
4	US	M	NA	James Kvarniq	James	Kvarniq	27JUN1970	4382 Gralyn Rd	92601
5	US	F	NA	Sandrina Stephano	Sandrina	Stephano	09JUL1975	6468 Cog Hill Ct	92601
6	BE	M	NA	Rent Van Lint	Rent	Van Lint	23DEC1945	Mispadstr 2	19001
7	ES	F	NA	Julián Escorihuela Monserrate	Julián	Escorihuela Monserrate	07AUG1975	Co. De Los Claveles 561	83001
8	FI	M	NA	Aki Ivonen	Aki	Ivonen	04DEC1935	Valimotie 5	34001
9	DE	F	NA	Cornelia Krahl	Cornelia	Krahl	27FEB1970	Kailstadterstr. 9	39401
10	US	F	NA	Kari Cornelia Krahl	Karen	Ballinger	18OCT1980	425 Bryant Estates Dr	92601
11	DE	F	NA	Elke Wallstab	Elke	Wallstab	16AUG1970	Carl-Zeiss-Str. 15	39401

Country	Country_Name	Population	Country_ID	Continent_ID	Country_Former_Name
AQ	Antarctica	.	11	90	
PR	Puerto Rico	.	72	91	
VI	Virgin Islands (U.S.)	.	78	91	
AW	Aruba	.	100	91	
BS	Bahamas	.	180	91	
BM	Bermuda	.	195	91	
BZ	Belize	.	227	91	British Honduras

DATA EXTRACTION

Evidence of data extraction

Order_ID	Order_Type	Employee_ID	Customer_ID	Order_Date	Delivery_Date
1230000029	3	99999999	21991	01JAN1998	01JAN1998
1230000126	3	99999999	48163	01JAN1998	01JAN1998
1230000360	3	99999999	[illegible]0531	01JAN1998	01JAN1998
1230000396	3	99999999	94039	01JAN1998	01JAN1998
1230000400	3	99999999	78431	01JAN1998	01JAN1998
1230000404	3	99999999	83582	01JAN1998	01JAN1998
1230000461	3	99999999	3290	01JAN1998	01JAN1998
1230000484	3	99999999	7435	01JAN1998	01JAN1998

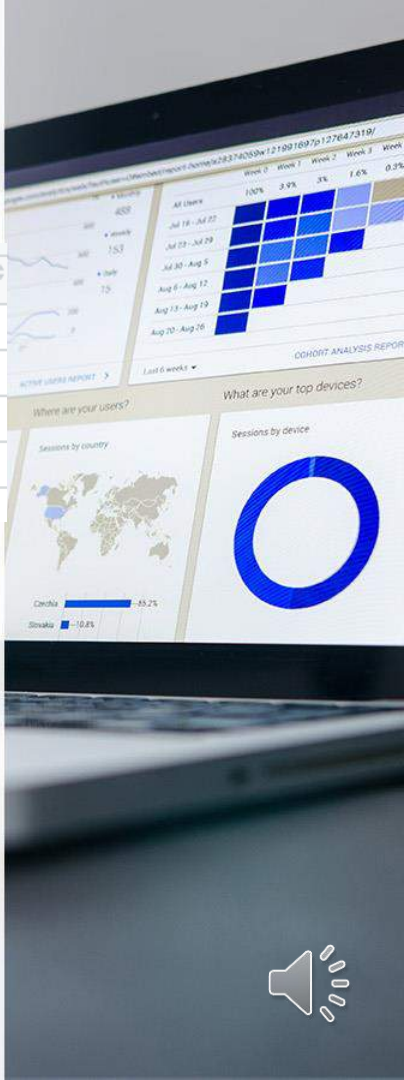
48163

	Order_ID	Order_Item_Num	Product_ID	Quantity	Total_Retail_Price	Discount
1	1230000029	1	240100100136	3	$547.80	.
2	1230000126	1	220101100034	2	$23.60	.
3	1230000360	1	220100100459	2	$173.20	.
4	1230000396	2	240100400095	4	$800.40	.
5	1230000400	1	240200100056	1	$40.80	.

	Product_Level	Product_Level_Name
1	1	Product
2	2	Product Group
3	3	Product Category
4	4	Product Line

Product_ID	Start_Date	End_Date	Unit_Cost_Price	Unit_Sales_Price	Factor
210100100001	01JAN1998	31DEC9999	$31.00	$67.80	1.00
210100100002	01JAN1998	22JAN1999	$55.15	$128.10	1.04
210100100002	23JAN1999	13FEB2000	$56.80	$131.90	1.03
210100100002	14FEB2000	06MAR2001	$57.90	$134.50	1.02
210100100002	07MAR2001	28MAR2002	$58.45	$135.80	1.01
210100100002	29MAR2002	31DEC9999	$58.45	$135.80	1.00
210100100003	01JAN1998	31DEC9999	$31.10	$69.90	1.00
210100100004	01JAN1998	31DEC9999	$27.85	$60.90	1.00

Product_ID	Product_Name	Supplier_ID	Product_Level	Product_Ref_ID
210000000000	Children	.	4	.
210100000000	Children Outdoors	.	3	210000000000
210100100000	Outdoor things, Kids	.	2	210100000000
210100100001	Boy's and Girl's Ski Pants with Braces	50	1	210100100000
210100100002	Children's Jacket	4742	1	210100100000
210100100003	Children's Jacket Sidney	50	1	210100100000
210100100004	Children's Rain Set	50	1	210100100000



DATA QUALITY CHECK

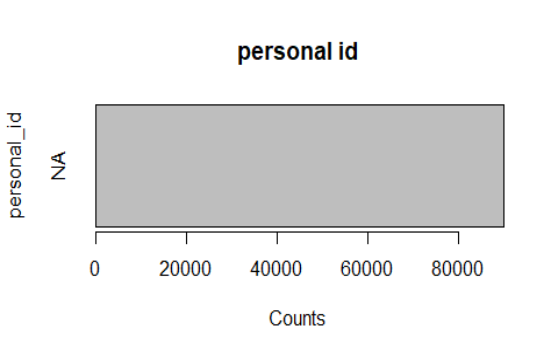
- Check and remove NA values

Untitled1* | customer_type | customer | country | order | product_level | product | price_list | order_item

Filter

	Customer_ID	Country	Gender	Personal_ID	Customer_Name	Customer_FirstName	Customer_LastName	Birth_Date	Customer_Address	Street_ID
1	1	FR	M	NA	Albert Collet	Albert	Collet	24NOV1940	Square Edouard VII 1	35001
2	2	ES	F	NA	Mercedes Martinez	Mercedes	Martinez	15JAN1955	Edificio 2	83001
3	3	IT	M	NA	Pier Egidio Boeris	Pier Egidio	Boeris	01JUL1970	Via M. Di Montescle 3	48001
4	4	US	M	NA	James Kvamiq	James	Kvamiq	27JUN1970	4302 Grelyn Rd	92601
5	5	US	F	NA	Sandrina Stephano	Sandrina	Stephano	09JUL1975	6468 Cog Hill Ct	92601
6	6	BE	M	NA	Rent Van Lint	Rent	Van Lint	23DEC1945	Mispadstr 2	15001
7	7	ES	F	NA	Julián Escolhuela Monserrate	Julián	Escolhuela Monserrate	07AUG1975	Co. De Los Cavalos 561	83001
8	8	FI	M	NA	Aki Ivonen	Aki	Ivonen	04DEC1965	Valimotie 5	34001
9	9	DE	F	NA	Cornelia Krahl	Cornelia	Krahl	27FEB1970	Kalstadterstr. 9	39401
10	10	US	F	NA	Karen Cornelia Krahl	Karen	Ballinger	18OCT1980	425 Bryant Estates Dr	92601
11	11	DE	F	NA	Elke Wellsteb	Elke	Wellsteb	16AUG1970	Carl-Zeiss-Str. 15	39401

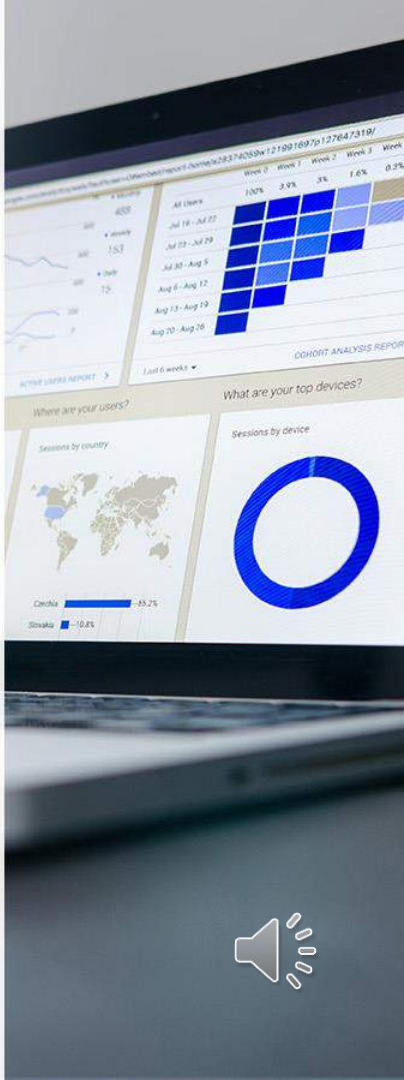
Showing 1 to 12 of 89,954 entries, 12 total columns



	Country	Country_ID
143	SO	800
144	ZA	801
145	ZW	818
146	NA	821
147	EH	831

```
> sum(is.na(country))
[1] 1
> which(is.na(country$Country_ID))
integer(0)
> which(is.na(country$Country))
[1] 146
> country <- na.omit(country)
> sum(is.na(country))
[1] 0
```

```
> sum(is.na(price_list))
[1] 0
> which(is.na(price_list$Product_ID))
integer(0)
> which(is.na(price_list$Start_Date))
integer(0)
> which(is.na(price_list$End_Date))
integer(0)
> which(is.na(price_list$Unit_Cost_Price))
integer(0)
> which(is.na(price_list$Unit_Sales_Price))
integer(0)
>
```



DATA QUALITY CHECK

- Evidence worked on special characters and crazy characters

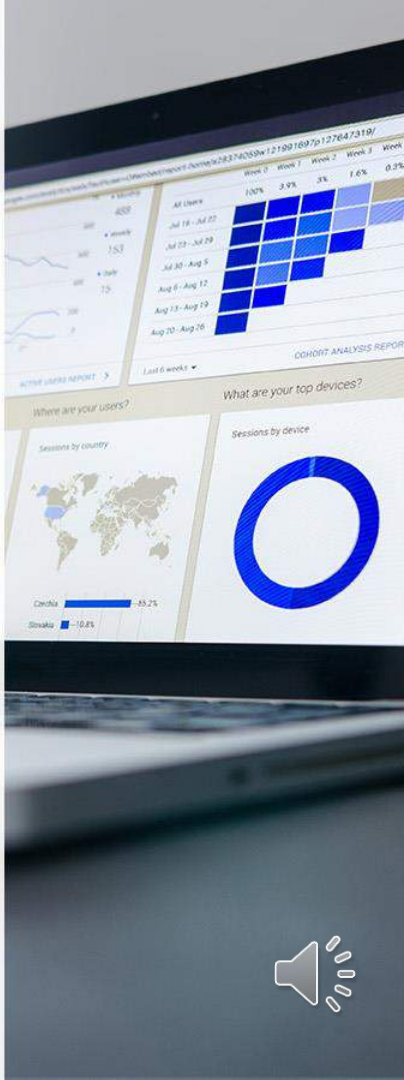
Customer_ID	Customer_Type_ID	Customer_Group_ID	Customer_Name	Countr
17648	1010	10	Kenneth Kastelberg	US
62552	1010	10	Milena Katia Libetti	IT
83806	1010	10	Julio Varela	ES
23021	1010	10	Cordula Gotzke	DE
26129	1010	10	B.T. Hargreaves	NL
85436	1010	10	Francisco Jose Falcon de Andres	ES
93090	1010	10	Jeanette Wallevik	NO

r_Group_ID	Customer_Name	Country	Customer_Type
10	Kenneth Kastelberg	US	Orion Club mem
10	Milena Katia Libetti	IT	Orion Club mem
10	Julio Varela	ES	Orion Club mem
10	Cordula Götzke	DE	Orion Club mem
10	B.T. Hargreaves	NL	Orion Club mem
10	Francisco José Falcón de Andrés	ES	Orion Club mem
10	Jeanette Wallevik	NO	Orion Club mem
10	Isabel Gordón	ES	Orion Club mem
10	Brendan Crawford	ZA	Orion Club mem

	Total_Retail_Price
3	$547.80
2	$23.60
2	$173.20
4	$800.40
1	$40.80

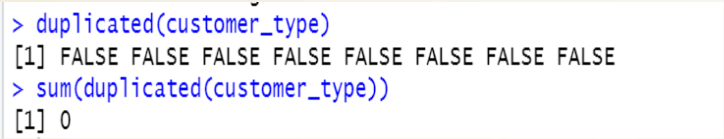
	Total_Retail_Price
3	547.80
2	23.60
2	173.20
4	800.40
1	40.80
3	60.60
1	12.90
1	51.20
1	133.10

```
> order_item$Total_Retail_Price<-  gsub("\\$", "", order_item$Total_Retail_Price)
> view(order_item)
```

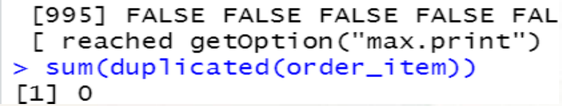


DATA QUALITY CHECK

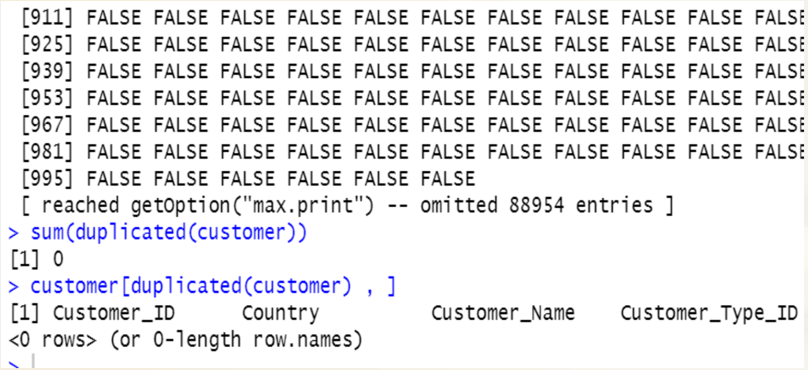
- Evidence for duplicates check



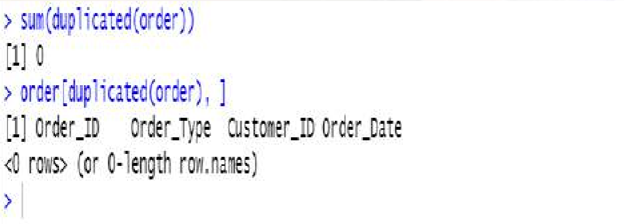
```
> duplicated(customer_type)
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> sum(duplicated(customer_type))
[1] 0
```



```
 [995] FALSE FALSE FALSE FALSE FAL
 [ reached getOption("max.print")
> sum(duplicated(order_item))
[1] 0
```

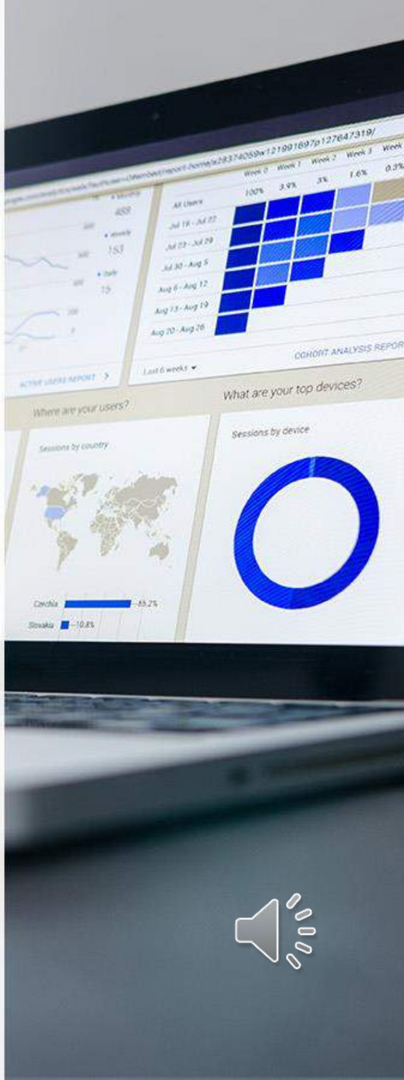


```
 [911] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [939] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [953] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [967] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [981] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [995] FALSE FALSE FALSE FALSE FALSE FALSE
 [ reached getOption("max.print") -- omitted 88954 entries ]
> sum(duplicated(customer))
[1] 0
> customer[duplicated(customer) , ]
[1] Customer_ID      Country          Customer_Name    Customer_Type_ID
<0 rows> (or 0-length row.names)
>
```



```
> sum(duplicated(order))
[1] 0
> order[duplicated(order), ]
[1] Order_ID    Order_Type  Customer_ID Order_Date
<0 rows> (or 0-length row.names)
>
```

```
> duplicated(customer)
```



DATA QUALITY CHECK

- Evidence for checking datatypes and converted

```
> glimpse(order)
Rows: 755,173
Columns: 4
$ Order_ID     <int> 1230000029, 1230000126, 1230000360, 1230000396, 1230000400, 12300004~
$ Order_Type   <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
$ Customer_ID  <int> 21991, 48163, 90531, 94039, 78431, 83582, 3290, 7435, 9587, 24730, 5~
$ Order_Date   <chr> "01JAN1998", "01JAN1998", "01JAN1998", "01JAN1998", "01JAN1998", "01~
> class(order$Order_Date)
[1] "character"
```

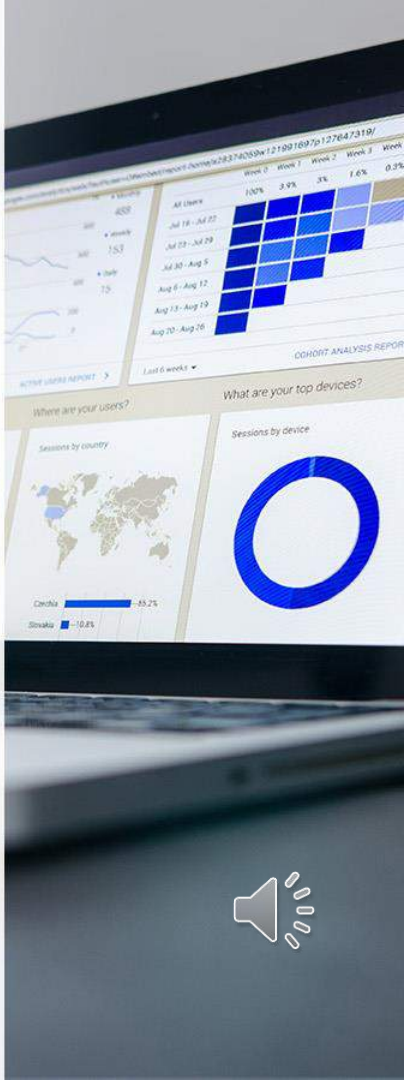
```
> order$Order_Date <-  as.Date(parse_date_time(order$Order_Date, c('dmy', 'ymd')))
> class(order$Order_Date)
[1] "Date"
>
```

```
> sapply(order_item, class)
          Order_ID     Order_Item_Num         Product_ID           Quantity
         "integer"          "integer"          "numeric"          "integer"
Total_Retail_Price
       "character"
> order_item$Total_Retail_Price <- as.numeric(as.character(order_item$Total_Retail_Price))
> class(order_item$Total_Retail_Price)
[1] "numeric"
>
```

```
> price_list = subset(price_l
> glimpse(price_list)
Rows: 20,733
Columns: 5
$ Product_ID        <dbl> 2101
$ Start_Date        <chr> "01J
$ End_Date          <chr> "31D
$ Unit_Cost_Price   <chr> "$31
$ Unit_Sales_Price  <chr> "$67
```

```
> glimpse(price_list)
Rows: 20,733
Columns: 5
$ Product_ID        <int64> 21
$ Start_Date        <date> 199
$ End_Date          <date> 999
$ Unit_Cost_Price   <dbl> 31.0
$ Unit_Sales_Price  <dbl> 67.8
```

```
> glimpse(customer)
Rows: 89,954
Columns: 4
$ Customer_ID       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, ~
$ Country           <chr> "FR", "ES", "IT", "US", "US", "BE", "ES", "FI", "DE", "US", "DE~
$ Customer_Name     <chr> "Albert Collet", "Mercedes Martínez", "Pier Egidio Boeris", "Ja~
$ Customer_Type_ID  <int> 2030, 1010, 1040, 1020, 2020, 1030, 1040, 1020, 2020, 1040, 104~
>
```



DATA VALIDATION

- Evidence for validation check

	Customer_Type_ID	Customer_Type	Customer_Group_ID	Customer_Group	Customer_ID	Country	Customer_Name
1	1010	Orion Club members inactive	10	Orion Club members	2	ES	Mercedes Martinez
2	1010	Orion Club members inactive	10	Orion Club members	14	FR	Albert Eulert
3	1010	Orion Club members inactive	10	Orion Club members	15	IT	Claudia Cambiaggi
4	1010	Orion Club members inactive	10	Orion Club members	32	AU	Gavin Graham
5	1010	Orion Club members inactive	10	Orion Club members	35	GB	Mike Merriott
6	1010	Orion Club members inactive	10	Orion Club members	55	US	Barner Matthews
7	1010	Orion Club members inactive	10	Orion Club members	62	US	Ihsan Robertson-Hector
8	1010	Orion Club members inactive	10	Orion Club members	87	FR	Marie-Helene Destombes
9	1010	Orion Club members inactive	10	Orion Club members	103	US	Judy Hicks
10	1010	Orion Club members inactive	10	Orion Club members	114	ES	Mónica Arévalo
11	1010	Orion Club members inactive	10	Orion Club members	118	BE	Annelies Paasonen
12	1010	Orion Club members inactive	10	Orion Club members	127	GB	Vivien Vager
13	1010	Orion Club members inactive	10	Orion Club members	135	US	Kertia Thomas
14	1010	Orion Club members inactive	10	Orion Club members	138	DE	Reinhild Von Bohlen
15	1010	Orion Club members inactive	10	Orion Club members	145	AT	Thomas Joksch
16	1010	Orion Club members inactive	10	Orion Club members	176	IT	Giuseppe Massimi
17	1010	Orion Club members inactive	10	Orion Club members	178	US	Cash Peoples
18	1010	Orion Club members inactive	10	Orion Club members	179	DE	Wiltraut Hessner
19	1010	Orion Club members inactive	10	Orion Club members	182	GB	Dierdre Andrews
20	1010	Orion Club members inactive	10	Orion Club members	202	MK	Dimitar Pavlovska
21	1010	Orion Club members inactive	10	Orion Club members	207	CH	Giuseppe Schärer

howing 1 to 2 f 89,954 entries total columns

	Customer_ID	Country	Gender	Personal_ID	Customer_Name	Customer_FirstName	Customer_LastName
1	1	FR	M	NA	Albert Collet	Albert	Collet
2	2	ES	F	NA	Mercedes Martinez	Mercedes	Martinez
3	3	IT	M	NA	Pier Egidio Boeris	Pier Egidio	Boeris
4	4	US	M	NA	James Kvarniq	James	Kvarniq
5	5	US	F	NA	Sandrina Stephano	Sandrina	Stephano
6	6	BE	M	NA	Rent Van Lint	Rent	Van Lint
7	7	ES	F	NA	Julián Escorihuela Monserrate	Julián	Escorihuela Monserrate
8	8	FI	M	NA	Aki Ivonen	Aki	Ivonen
9	9	DE	F	NA	Cornelia Krahl	Cornelia	Krahl
10	10	US	F	NA	Karl Cornelia Krahl	Karen	Ballinger
11	11	DE	F	NA	Elke Wallstab	Elke	Wallstab
12	12	US	M	NA	David Black	David	Black

Show to 12 of 89,954 entries 2 total columns

DATA VALIDATION

Evidence for data validation before and after

Filter

	Country	Country_ID
1	AQ	11
2	PR	72
3	VI	78
4	AW	100
5	BS	180
6	BM	195
7	BZ	227
8	VG	231
9	CA	260
10	KY	268
11	CR	295
12	CU	300

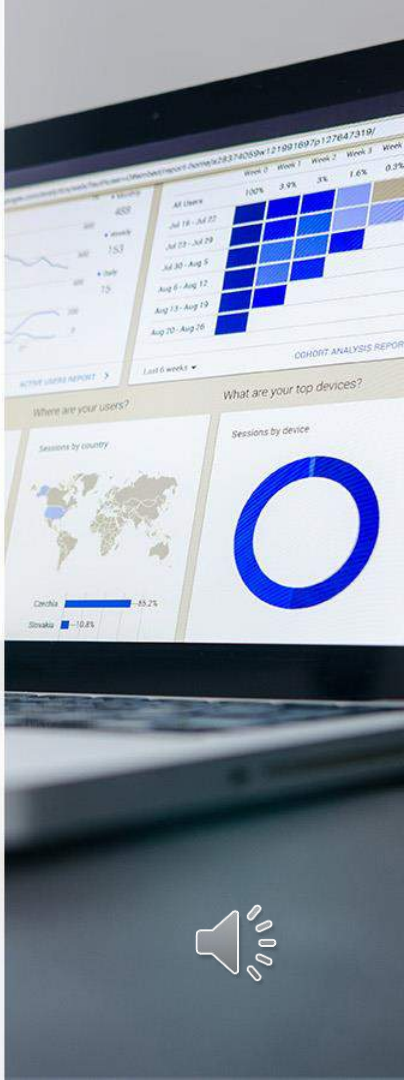
Showing 1 to 13 of 236 entries, 2 total columns

ImportOrionData.R* customer produ

Filter

	Country	Country_ID
1	AQ	11
2	PR	72
3	VI	78
4	AW	100
5	BS	180
6	BM	195
7	BZ	227
8	VG	231
9	CA	260
10	KY	268
11	CR	295
12	CU	300

Showing 1 to 13 of 235 entries, 2 total columns



DATA INTEGRATION

- Data integration involves combining different data into standardized formats and to be stores at database as data marts which give insights for informed decision making for business.

Here we merged customer dataset and customer type dataset by using common column customer type Id. In such a product data set combined with product level dataset by using product level.

Evidence of data integration

```
#merge customer_type with customer df
customer <- merge(customer,customer_type, by= 'Customer_Type_ID')
View(customer)
```

```
#merge product_type with product df
product <- merge(product, product_level, by ='Product_Level')
```

DATA LOADING

Data is loading into operational data for quick analysis where our process can access the data in order to bring reports for required business question. By using putty we access the hive and create database over their to create and load the data into database for easy access.

```
INFO  : OK
+----------------------+
|    database_name     |
+----------------------+
| default              |
| foodmart             |
| information_schema   |
| orion_star           |
| sys                  |
+----------------------+
5 rows selected (0.394 seconds)
```

```
INFO  : OK
+----------------------+------------+----------+
|      col_name        | data_type  | comment  |
+----------------------+------------+----------+
| order_id             | int        |          |
| customer_id          | int        |          |
| country_id           | int        |          |
| product_id           | bigint     |          |
| timeid               | int        |          |
| quantity             | int        |          |
| total_retail_price   | double     |          |
| unit_cost_price      | double     |          |
| unit_sales_price     | double     |          |
| totalamountsales     | double     |          |
+----------------------+------------+----------+
```

```
INFO  : OK
+----------------+------------+----------+
|    col_name    | data_type  | comment  |
+----------------+------------+----------+
| customer_id    | int        |          |
| customer_name  | string     |          |
| customer_type  | string     |          |
| country        | string     |          |
+----------------+------------+----------+
```

DATA LOADING

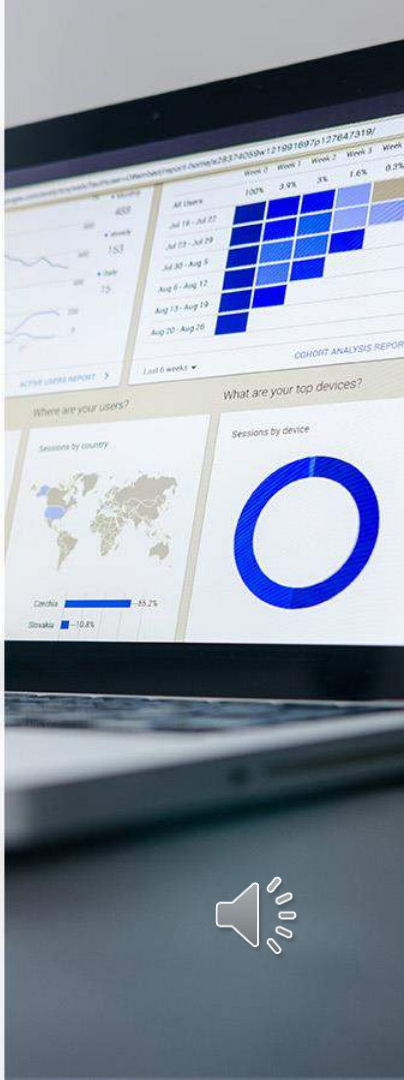
Evidence for creation and count for the table

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Select Count(*) from  dimcountry;
INFO  : Compiling command(queryId=hive_20230425193014_da4156d5-4909-48f1-ba23-d1715e
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_c0, type:bigin
INFO  : Completed compiling command(queryId=hive_20230425193014_da4156d5-4909-48f1-b
INFO  : Executing command(queryId=hive_20230425193014_da4156d5-4909-48f1-ba23-d1715e
INFO  : Completed executing command(queryId=hive_20230425193014_da4156d5-4909-48f1-b
INFO  : OK
+------+
| _c0  |
+------+
| 235  |
```

```
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2> Select Count(*) from  dimcustomer;
INFO  : Compiling command(queryId=hive_20230425193135_555ef90f-01aa-4bc6-bba4-5f5df
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:_c0, type:bigi
INFO  : Completed compiling command(queryId=hive_20230425193135_555ef90f-01aa-4bc6-
INFO  : Executing command(queryId=hive_20230425193135_555ef90f-01aa-4bc6-bba4-5f5df
INFO  : Completed executing command(queryId=hive_20230425193135_555ef90f-01aa-4bc6-
INFO  : OK
+--------+
|  _c0   |
+--------+
| 89954  |
+--------+
```

```
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSch
e:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230331091
INFO  : Executing command(queryId=hive_20230331091713_7d8648
INFO  : Completed executing command(queryId=hive_20230331091
INFO  : OK
+------------------------+---------------------+
| dimcountry.country_id  | dimcountry.country  |
+------------------------+---------------------+
+------------------------+---------------------+
No rows selected (0.615 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```

```
INFO  : Executing command(queryId=hive_20230331091044_29912b4f-a110-4a66-8321-6bb21f300b88): SELECT * FROM DIMC
INFO  : Completed executing command(queryId=hive_20230331091044_29912b4f-a110-4a66-8321-6bb21f300b88); Time take
INFO  : OK
+-------------------------+--------------------------+--------------------------+---------------------+
| dimcustomer.customer_id | dimcustomer.customer_name | dimcustomer.customer_type | dimcustomer.country |
+-------------------------+--------------------------+--------------------------+---------------------+
+-------------------------+--------------------------+--------------------------+---------------------+
No rows selected (1.249 seconds)
0: jdbc:hive2://sandbox-hdp.hortonworks.com:2>
```



DATA LOADING

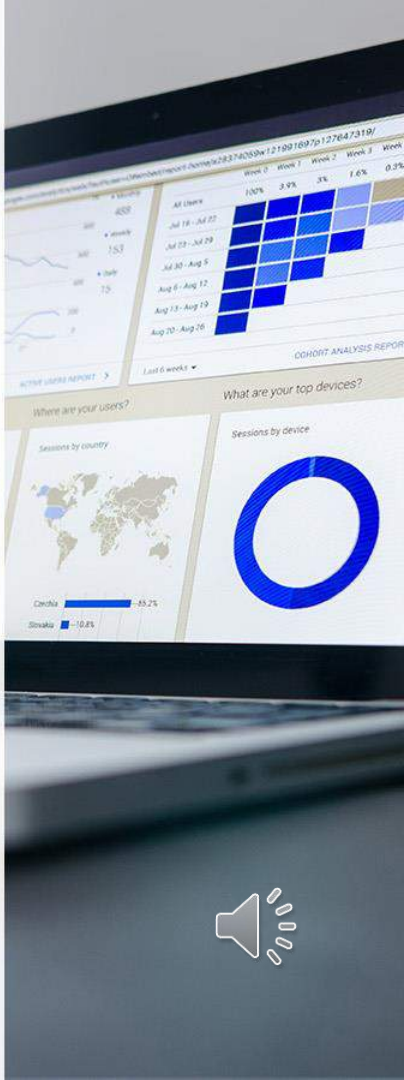
Evidence of datamarts creation

```
INFO  : OK
+-------------+------------+----------+
|  col_name   | data_type  | comment  |
+-------------+------------+----------+
| timeid      | int        |          |
| order_date  | string     |          |
| day         | int        |          |
| month       | int        |          |
| year        | int        |          |
+-------------+------------+----------+
5 rows selected (0.155 seconds)
```

```
INFO  : OK
+---------------------+------------+----------+
|      col_name       | data_type  | comment  |
+---------------------+------------+----------+
| product_id          | bigint     |          |
| product_name        | string     |          |
| product_level_name  | string     |          |
+---------------------+------------+----------+
3 rows selected (0.15 seconds)
```

```
INFO  : OK
+-----------------+
|    tab_name     |
+-----------------+
| dimcountry      |
| dimcustomer     |
| dimproduct      |
| dimtime         |
| orderfacttable  |
+-----------------+
```

```
INFO  : OK
+-------------+------------+----------+
|  col_name   | data_type  | comment  |
+-------------+------------+----------+
| country_id  | int        |          |
| country     | string     |          |
+-------------+------------+----------+
```



STAR SCHEMA

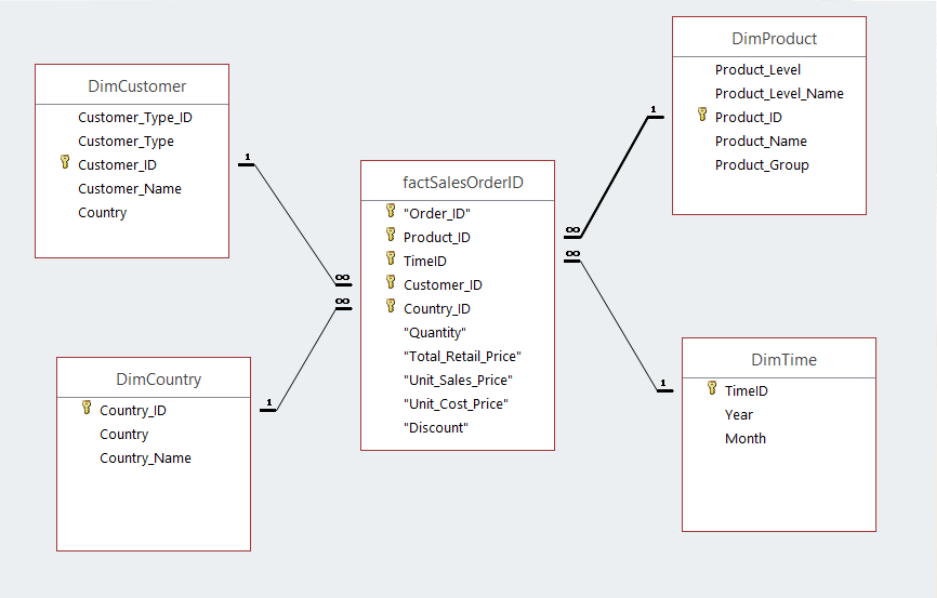
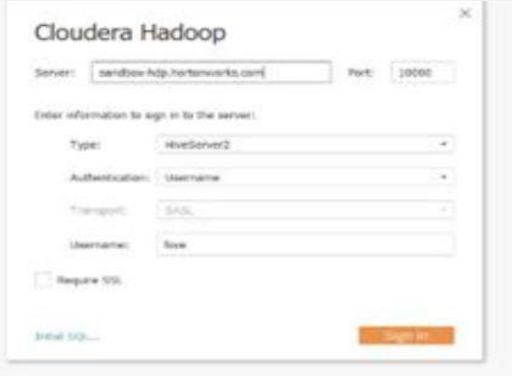


Figure: Data Dimensional Model

- Star Schema is also called multi dimensional data model which is organized with dimensions and fact data marts.
- It is developed for querying large data sets working faster and easier.
- This is used to denormalized business data into single fact table and It connects with other multiple dimension tables as shown in the figure.
- Great for basic queries since they rely less on joins when retrieving data than normalized models like snowflake schemas.

DATA VISULALIZATION

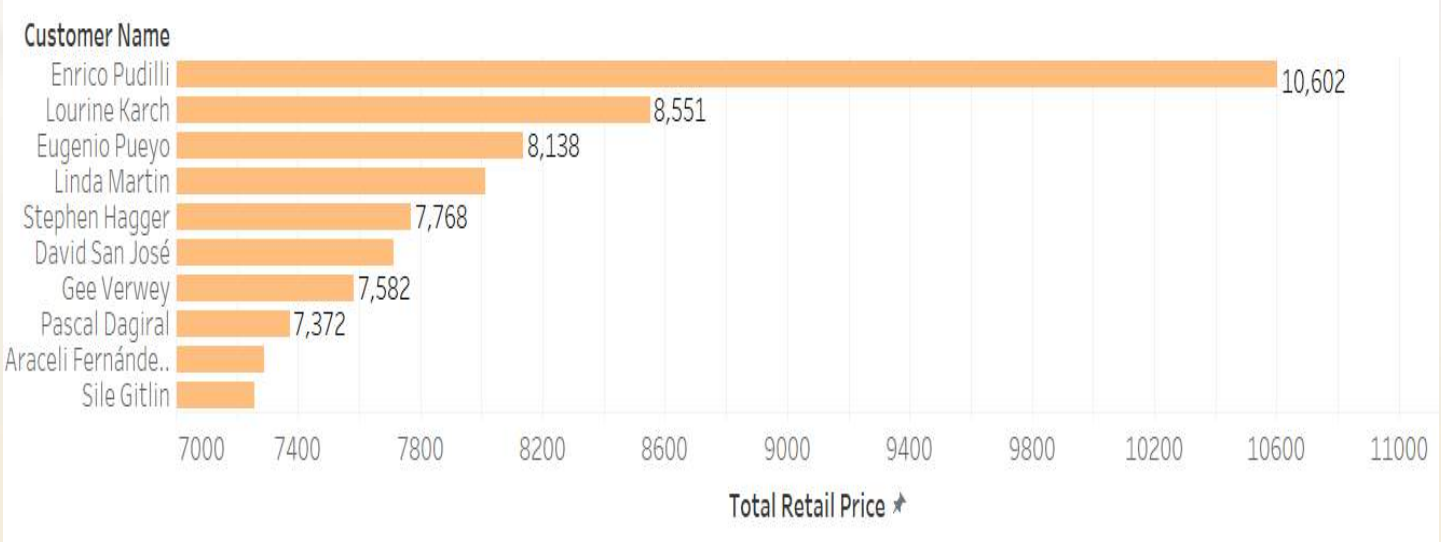
- For all provided business questions I got reports in graphical and tabular form by using tableau tool.
- The dimensional data of Hive transfer to tableau, the connection established by using cloudera.
- To avoid errors by downloading and uploading file we use integrated tools.
- Finally the dashboard created from data which can understand easily. We can visualize the reports and make informed data driven decisions by organization.



Q1) Top 10 Customers

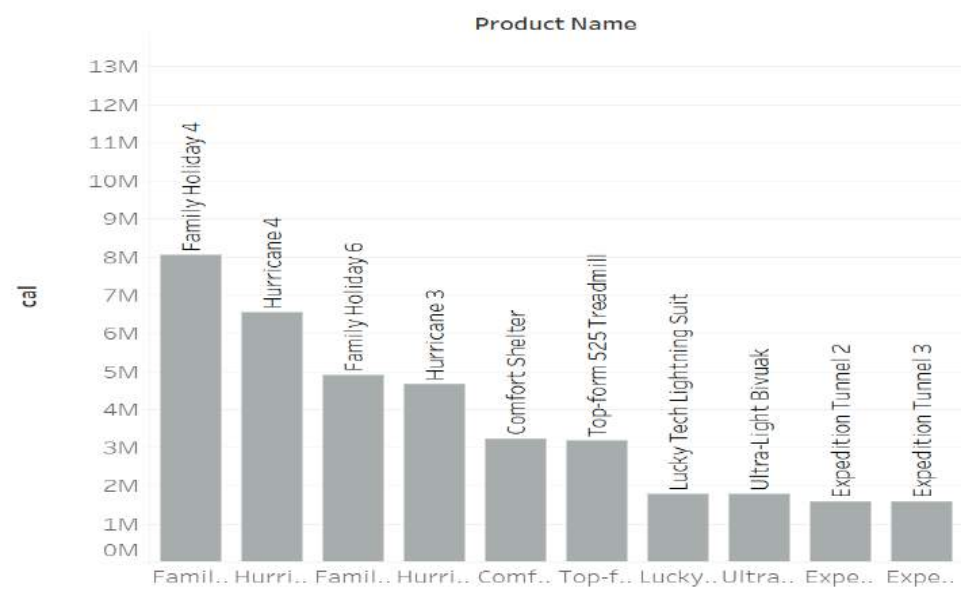
- Here we can see top 10 customers by sales value in last month which is 2002 DEC

Customer Name	
Enrico Pudilli	10,602
Lourine Karch	8,551
Eugenio Pueyo	8,138
Linda Martin	8,014
Stephen Hagger	7,768
David San José	7,711
Gee Verwey	7,582
Pascal Dagiral	7,372
Araceli Fernánde..	7,290
Sile Gitlin	7,258



Q2)Most Profitable Product

Most profitable product of 2002 Family Holiday 4



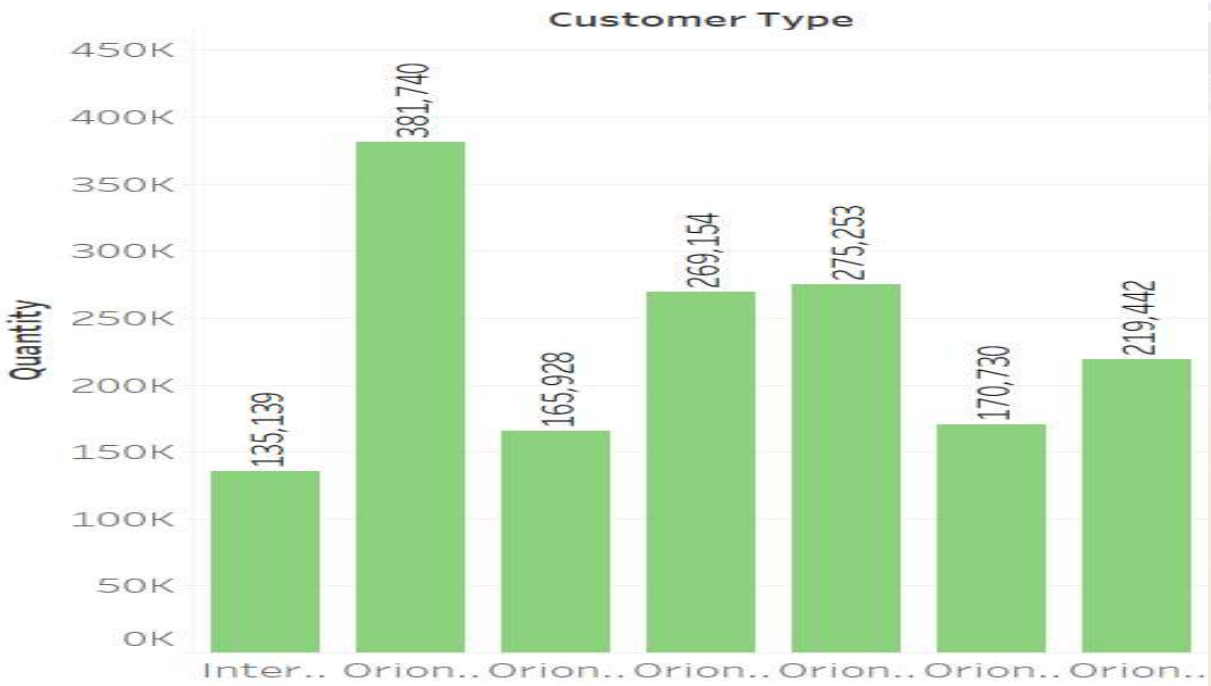
Product Name	cal
Family Holiday 4	8,075,120
Hurricane 4	6,561,536
Family Holiday 6	4,915,560
Hurricane 3	4,672,749
Comfort Shelter	3,245,130
Top-form 525 Tr..	3,202,414
Lucky Tech Light..	1,788,349
Ultra-Light Bivuak	1,787,801
Expedition Tunn..	1,609,988
Expedition Tunn..	1,603,063

cal

2M 8M

Q3)Total Products Purchased By Customer Type

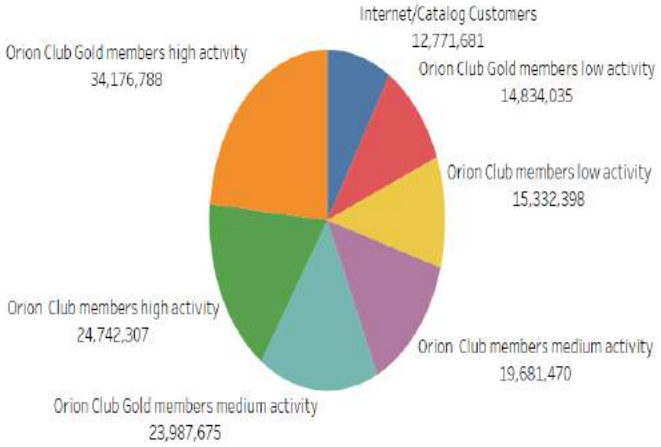
Overall products purchased in 2002 by customer type



Customer Type	
Internet/Catalog Customers	135,139
Orion Club Gold members high activity	381,740
Orion Club Gold members low activity	165,928
Orion Club Gold members medium activity	269,154
Orion Club members high activity	275,253
Orion Club members low activity	170,730
Orion Club members medium activity	219,442

Q4)Revenue Generated By Customer Type

Least amount generated in 2002 by customer type



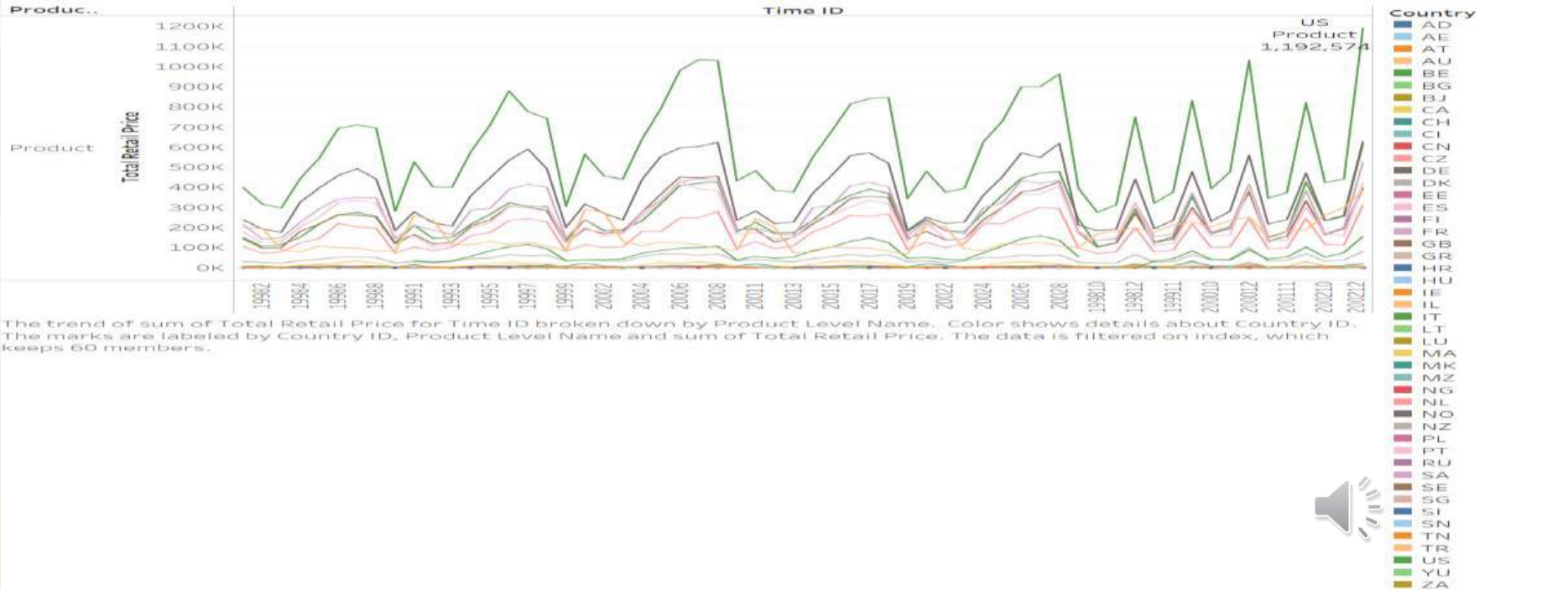
Total Retail Price
145,526,353

Customer Type
- Internet/Catalog Customers
- Orion Club Gold members low activity
- Orion Club members low activity
- Orion Club members medium activity
- Orion Club Gold members medium activity
- Orion Club members high activity
- Orion Club Gold members high activity

Customer Type	
Internet/Catalog Customers	12,771,681
Orion Club Gold members low activity	14,834,035
Orion Club members low activity	15,332,398
Orion Club members medium activity	19,681,470
Orion Club Gold members medium activity	23,987,675
Orion Club members high activity	24,742,307
Orion Club Gold members high activity	34,176,788

Q5) Total Sales By Country By Product Group

Graphical report for total sales by country by product group and here I used only country code instead of country name to view report properly and fit in a window



Q5. (Tabular Form)

Tabular form report for total sales by country by product group and here cropped the image for clear view

Product	Countr..	19981	19982	19983	19984	19985	19986	19987	19988	19989	19991	19992	19993	19994	19995	19996	19997	19998	19999	20001	2000
Product	AD									146				29		1,141					
	AE								253		101			1,542							
	AT	3,261	1,374	743	8,163	4,853	7,288	8,673	8,768	1,277	2,268	2,100	992	7,744	2,129	4,256	4,180	3,492	947	1,913	1,38
	AU	200,948	196,621	88,852	78,095	108,916	100,644	98,100	82,845	82,073	259,052	238,124	112,826	114,826	130,227	115,505	126,044	107,969	83,693	289,185	275,97
	BE										34,687	27,758	34,493	56,075	83,949	104,522	115,706	95,344	35,162	39,723	39,77
	BG			141		5	35		657	13				305			94	186	11		34
	BJ					32		588	442												
	CA	7,609	4,883	2,975	12,534	17,867	25,442	34,014	23,795	4,037	6,435	6,163	3,836	10,743	19,113	26,759	23,435	21,278	3,829	6,221	3,14
	CH	8,932	9,644	1,994	4,354	3,391	3,930	7,674	10,228	3,125	15,689	2,778	1,272	2,728	9,829	5,764	4,500	13,632	8,396	22,870	8,71
	CI		187					45	457				202			284	427		73		
	CN						68		602				109				1,099				
	CZ							491			359							117			
	DE	238,004	192,765	176,360	327,189	396,872	459,049	492,358	441,489	185,461	278,531	222,730	207,602	380,142	446,890	535,484	590,242	492,116	200,357	316,840	271,89
	DK	30,818	27,306	23,694	35,342	41,583	51,718	54,832	49,175	24,185	31,518	25,783	31,117	44,364	47,517	65,648	58,493	61,814	34,283	39,028	34,15
	EE	109	65			706	639				305		94			275	171	55		345	19
	ES	177,701	122,709	131,072	204,342	252,679	323,828	334,705	321,702	155,514	158,689	104,191	122,898	197,577	224,759	303,619	300,853	308,969	122,081	201,418	157,69
	FI	1,060	1,133	363	1,542	1,883	1,769	1,830	1,520	217	1,250	601	335	1,225	1,953	2,469	2,338	3,398	701	1,296	86
	FR	214,484	141,241	149,135	225,322	292,414	341,096	348,913	347,676	144,038	209,742	187,089	165,819	286,216	294,410	389,418	415,803	401,046	148,261	239,238	189,69
	GB	142,321	97,388	100,032	181,331	217,004	262,467	262,754	251,259	118,932	165,867	117,749	126,339	207,764	238,634	305,559	298,922	285,123	122,739	193,521	170,48
	GR	58	94	145		61	50	1,860	132		142	107	139	90	402	575	1,658	150		307	2
	HR		170	31	233	171	452	49	1,514		125	235		714	367	271	177	35			5
	HU	80	153	200	231	344	843	171	43		178		36	27	126	300	116	527		63	
	IE		39			128	46						24		402	177		50		47	
	IL	63			60	25		292	229	212		330	226	176	15	158	465	442			14
	IT	149,424	105,910	114,674	151,825	217,863	264,411	275,414	255,921	122,301	212,079	145,158	151,527	216,749	265,486	321,937	305,923	303,572	134,188	238,962	183,67
	LT	94									127	228			113		136	312		133	
	LU	627		445	213	876	577	1,771	754	391	336	407	1,183	401	1,194	375	1,327	374		79	15
	MA	59									252										
	MK					379								394	444		400				
	MZ								158								390	257			
	NG																			107	
	NL	105,965	72,012	78,703	123,130	144,891	220,407	202,992	198,018	72,684	103,665	84,617	102,266	158,973	174,909	235,118	244,129	227,333	81,455	114,418	101,15
	NO	3,986	2,008	3,106	3,799	6,953	8,984	4,105	6,263	1,738	2,801	2,505	1,540	4,512	9,364	4,957	10,777	3,907	1,592	2,121	2,69
	NZ					962	227	1,304			631	93			1,055			225			
	PL				277									736			267			47	2
	PT	6,996	5,802	3,977	5,240	15,690	14,109	6,851	10,408	909	6,581	3,656	3,844	7,981	8,476	13,001	15,489	12,135	2,844	6,673	4,38
	RU		323		252				647						1,202	706	512		390		
	SA	201	130		1,650		1,626	504			823							1,658			
	SE	4,316	3,394	1,182	2,901	5,928	7,836	7,116	2,926	589	2,664	3,744	2,543	3,910	7,780	3,544	8,865	5,080	959	4,004	3,46
	SG					123				35						264					
	SI	1,132	1,329		243	1,336	797		732	266	360	73	180	1,050	2,146	733	1,536	1,037	1,448	93	
	SN	182	452						398							226	266		97		
	TN				218						226					278					
	TR	2,318	3,362	4,150	5,610	5,336	8,492	7,254	5,389	842	5,769	2,405	1,462	5,023	9,662	7,857	7,701	10,423	1,489	2,144	2,26
	US	398,080	317,385	296,724	443,819	545,367	693,202	711,253	694,974	280,810	527,102	402,031	400,236	578,985	709,408	879,604	776,260	742,939	305,548	564,677	457,24
	YU	30											328								
	ZA	547	654	802	1,768	1,342	3,116	1,207	1,335	243	789	161	329	4,043	2,137	3,468	2,789	1,449	1,332	836	20

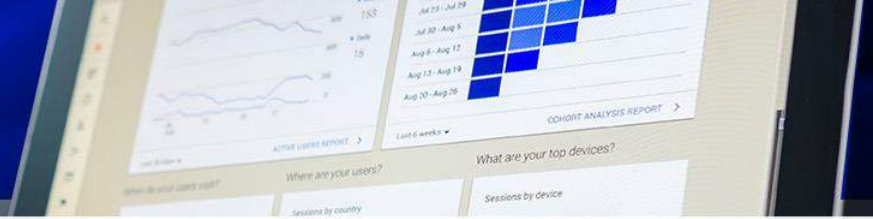
RECOMMENDATIONS

From the insights the business has to focus on few points for next sales

- By top customers we also know that the service and the products make available to all customers and we focus on their interests to bring to other customers.
- From most profitable product we can upgrade other products to increase the sales from that products.
- Overall product purchased by customer type is orion club gold members high activity from these business can offer better discounts to increase sales by other types.
- Least amount generated by customer type is Internet/catalogue customers, company can invest in this type to market to reach people.
- US has more sales on products, so company can start production and maintain inventory in following top countries for fast services

Overall, to increase the revenue of company, invest in the areas for marketing to reach customer about products and upgrade products which has less sales.

REFERENCES



- *What is Star Schema?* (n.d.). Databricks. https://www.databricks.com/glossary/star-schema#:~:text=What%20is%20a%20star%20schema
- Tehreem Naeem. (2019, October 3). *Data Warehouse Concepts: Kimball vs. Inmon Approach | Astera*. Astera. https://www.astera.com/type/blog/data-warehouse-concepts/
- Burns, E. (2021, January 14). *Data Cleaning in R Made Simple*. Medium. https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17
- *Data Cleaning in R*. (2022, June 25). GeeksforGeeks. https://www.geeksforgeeks.org/data-cleaning-in-r/
- *regex - Remove all special characters from a string in R?* (n.d.). Stack Overflow. https://stackoverflow.com/questions/10294284/remove-all-special-characters-from-a-string-in-r
- w3schools. (2019). *SQL INNER JOIN Keyword*. W3schools.com. https://www.w3schools.com/sql/sql_join_inner.asp
- Data Warehouse Design - javatpoint. (n.d.). *Www.javatpoint.com*. https://www.javatpoint.com/data-warehouse-design
- SHU HALLAM LEARNING MATERIALS

APPENDIX

Full table report for business question 5.

R Screenshots

```
> write.csv(price_list, "C:\\Users\\Mahendra\\OneDrive\\Desktop\\Re assesment  ADMP\\Datasets
\\price_list.csv", row.names=FALSE)
> view(price_list)
```

```
> order_item%>% filter(is.na(Order_ID))
[1] Order_ID            Order_Item_Num      Product_ID          Quantity
[5] Total_Retail_Price
<0 rows> (or 0-length row.names)
> order_item%>% filter(is.na(Product_ID))
[1] Order_ID            Order_Item_Num      Product_ID          Quantity
[5] Total_Retail_Price
<0 rows> (or 0-length row.names)
> order_item%>% filter(is.na(Quantity))
[1] Order_ID            Order_Item_Num      Product_ID          Quantity
[5] Total_Retail_Price
<0 rows> (or 0-length row.names)
> order_item%>% filter(is.na(Total_Retail_Price))
[1] Order_ID            Order_Item_Num      Product_ID          Quantity
[5] Total_Retail_Price
<0 rows> (or 0-length row.names)
> order_item%>% filter(is.na(Order_Item_Num))
[1] Order_ID            Order_Item_Num      Product_ID          Quantity
[5] Total_Retail_Price
<0 rows> (or 0-length row.names)
>
> customer %>% select(Customer_ID, Country,Customer_Name,Customer_Type_ID) %>%
+   filter(!complete.cases(.))
[1] Customer_ID       Country           Customer_Name     Customer_Type_ID
<0 rows> (or 0-length row.names)
```


ACTIVE USERS REPORT
COHORT ANALYSIS REPORT
Where are your users?
What are your top devices?
Sessions by country
Sessions by device


THANK YOU!