Case Study Group 5

SMIT PANCHAL – B00828070

MEETA CHANCHLANI – B00835734

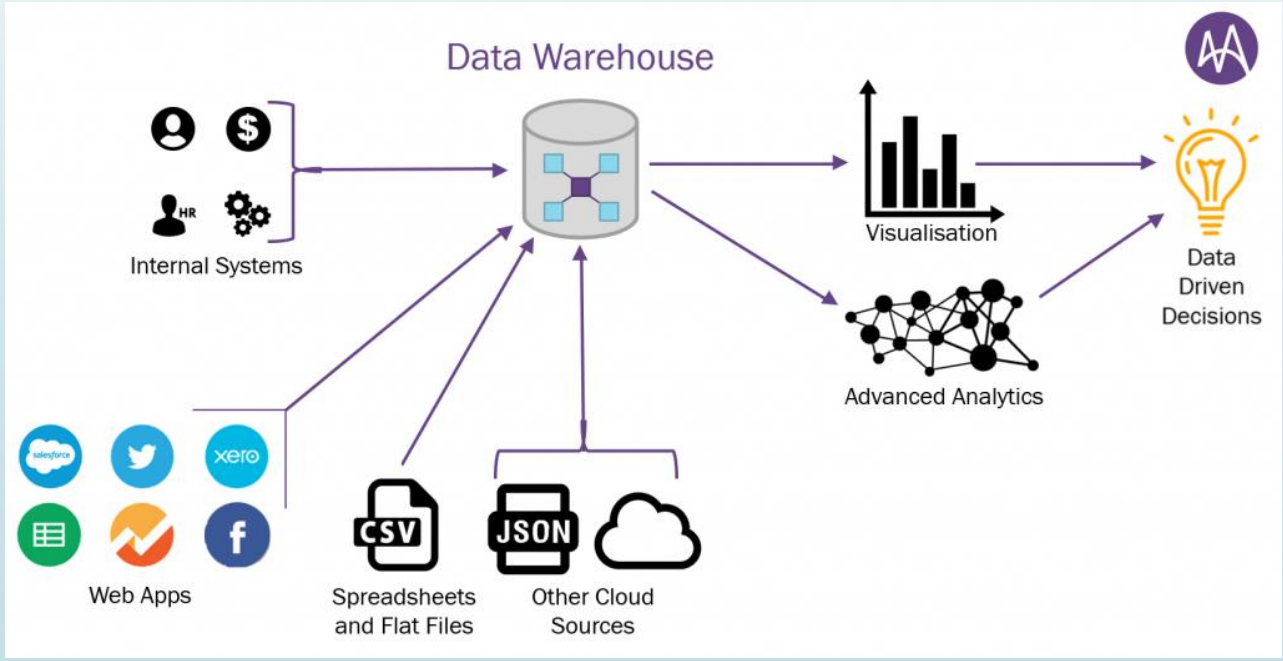
VIVEK SAKARIYA – B00848519

AGENDA

- Data warehousing
- ETL
- Dataset exploration
- Data cleaning
- Multi-dimensional Schema Development
- Python Script
- Business Intelligence
- Cognos BI

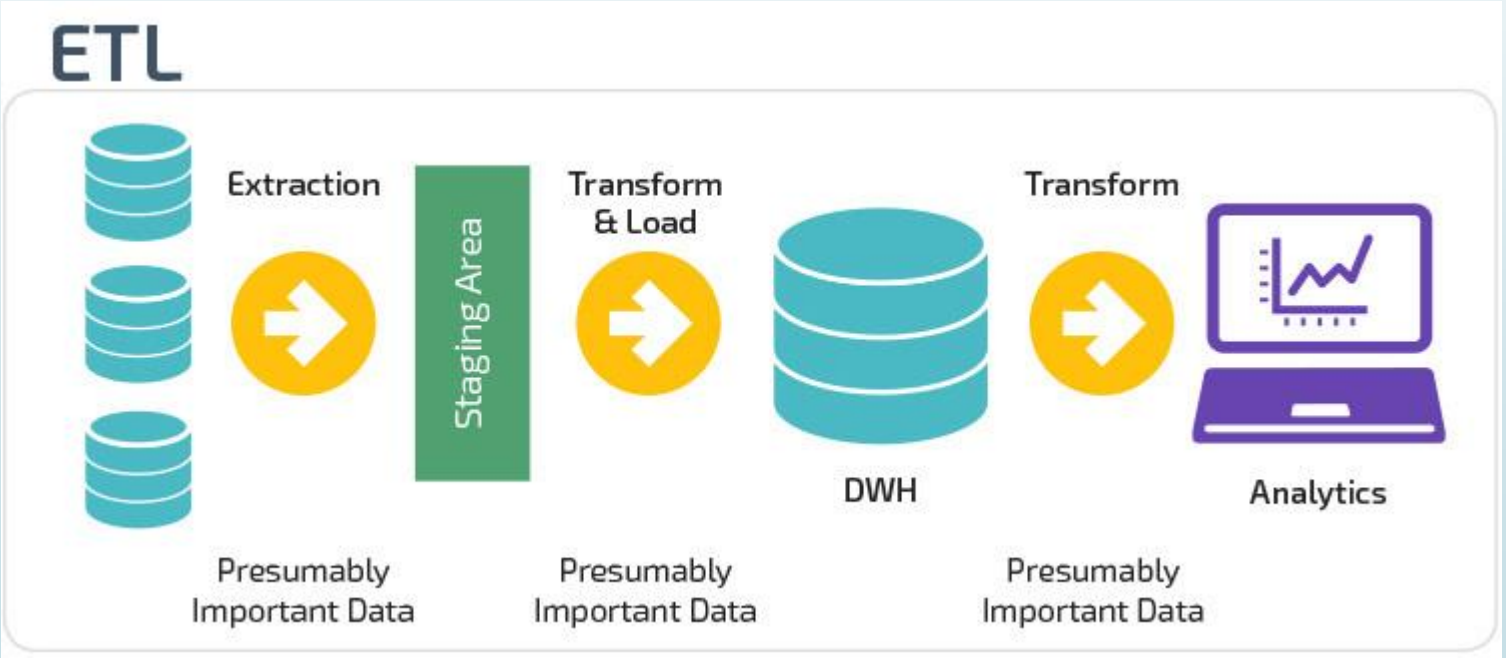
Data Warehousing

- It is process of constructing and using a data warehouse.
- A data warehouse is constructed by integrating different sources and managing it in a meaningful way.
- It generally involves 3 steps such as data cleaning, data integration, and data consolidations.



ETL

- Extract – collecting data from various sources
- Transform – convert the data into understandable form
- Load – store the transformed data on the target system

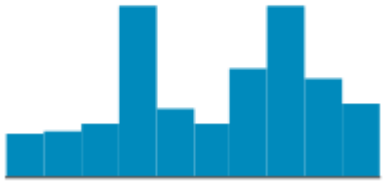


Dataset Exploration

- There are 25 fields in the sales_data_sample.csv
- The data is related to a specific sales data.
- There are various fields which can define the relationships and further be used for visualization.
- What is Sales dataset all about?
  - The dataset is all about an Automobile Sales.
  - It was inspired for retail analytics.
  - It was originally written by Maria Carina.
  - There are total 25 columns.
  - There are total 2823 rows.

- General observation of the Data?
  - The data is from year 2003 to 2005.
  - There are total 109 unique products.
  - There are total 19 unique countries.
  - There are 92 unique customers.
  - There are 307 unique orders.
  - There are total 4 territories.

Data Cleaning

ORDERDATE

5Jan03 30May05
2/24/2003 0:00
5/7/2003 0:00
7/1/2003 0:00

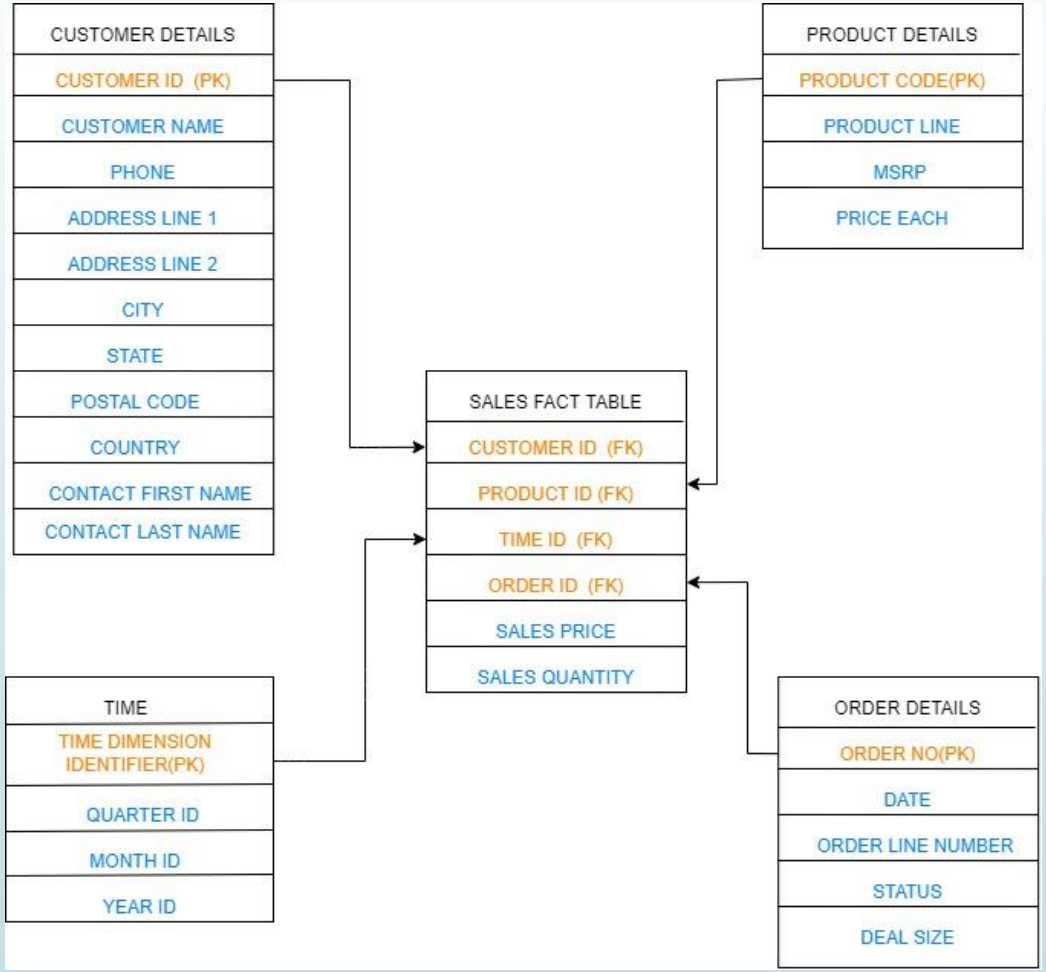
ADDRESSLINE2	
[null]	89%
Level 3	2%
Other (8)	9%

STATE		POSTALCODE	
[null]	53%	28034	9%
CA	15%	97562	7%
Other (15)	33%	Other (72)	84%
NY		10022	
		5020	
NSW		2067	

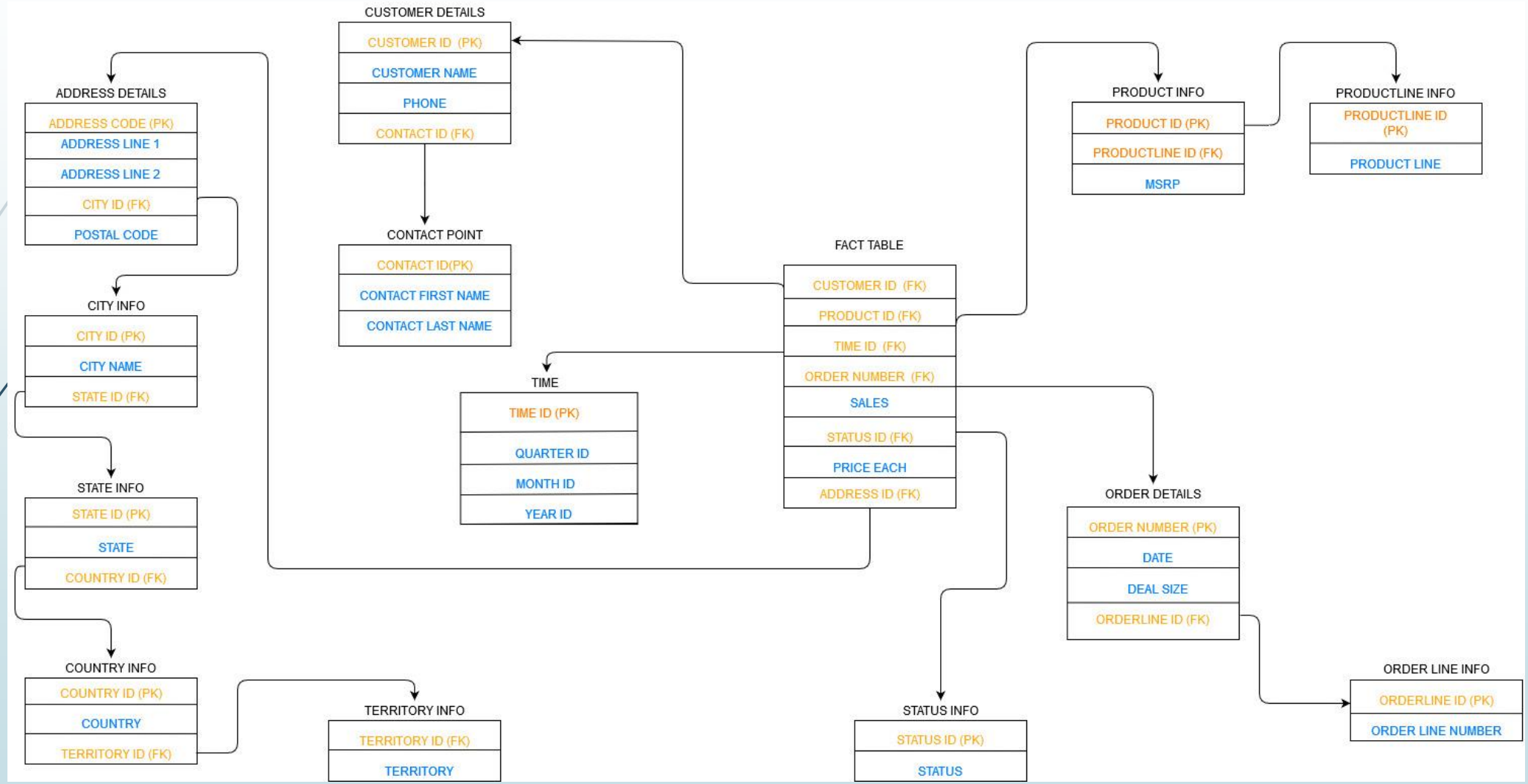
TERRITORY	
EMEA	50%
NA	38%
Other (2)	12%
NA	
EMEA	
APAC	

Multi-Dimensional Schema

- **Initial Schema for design (Star Schema): Option 1**



Subsequent Schema (Snowflakes Schema): Option 2



Python Script

```
import pandas as pd
```

Importing pandas library to work with dataframe

```
readcsv['ORDERDATE'] = readcsv['ORDERDATE'].str.split(" ",1,expand = True)
df = readcsv.fillna({
'ADDRESSLINE2' : 'NA',
'STATE' : 'NA',
'POSTALCODE' : 'NA',
'TERRITORY' : 'NA'

})

mywrite = pd.DataFrame(df, columns=['ORDERNUMBER','QUANTITYORDERED','PRICEEACH',
                                    'ORDERLINENUMBER','SALES','ORDERDATE','STATUS',
                                    'QTR_ID','MONTH_ID','YEAR_ID','PRODUCTLINE',
                                    'MSRP',
                                    'PRODUCTCODE','CUSTOMERNAME','PHONE',
                                    'ADDRESSLINE1',
                                    'ADDRESSLINE2','CITY','STATE','POSTALCODE',
                                    'COUNTRY',
                                    'TERRITORY','CONTACTLASTNAME',
                                    'CONTACTFIRSTNAME',
                                    'DEALSIZE'])

# Writing all rows to csv file
mywrite.to_csv('AllCleanedDataFile.csv', index=False)
print("All Cleaned Data File has been created")


newreadcsv = pd.read_csv('AllCleanedDataFile.csv', encoding="latin1")

orderDetails = pd.DataFrame(newreadcsv, columns=['ORDERNUMBER','ORDERDATE', 'DEALSIZE'])

od = orderDetails.drop_duplicates(subset='ORDERNUMBER',keep="first")
```

Extracting the fields which are needed for making the schema

```
for d in dfstatlist:
    for o in statlist:
        if o[0] == d[1]:
            statList.append(d)



statusdf = pd.DataFrame(statList, columns=['STATUSID', 'STATUS'])

statusdf.to_csv("STATUS.csv", index=False)
print("STATUS.csv has been created")

# PRODUCT INFO

productDetails = pd.DataFrame(newreadcsv, columns=['PRODUCTCODE', 'MSRP'])

pro = productDetails.drop_duplicates(subset='PRODUCTCODE',keep="first")
pro.insert(0, 'PRODUCTLINEID', range(80000, 80000 + len(pro)))



proList = []

dfprolist = pro.values.tolist()

productlist = productDetails.values.tolist()
productlist.sort()


#for fo in orderlist:
#    print(fo)

for d in dfprolist:
```

Separation of various dimension tables according to the snowflakes schema

```
print("PRODUCT.csv has been created")

# CUSTOMER DETAILS


custDetails = pd.DataFrame(newreadcsv, columns=['CUSTOMERNAME','PHONE'])

tem = custDetails.drop_duplicates(subset='CUSTOMERNAME',keep="first")

tem.insert(0, 'CUSTOMERID', range(82825, 82825 + len(tem)))
tem.insert(0, 'CONTACTID', range(94130, 94130 + len(tem)))


finalList = []

dflist = tem.values.tolist()


custlist = custDetails.values.tolist()
custlist.sort()


for d in dflist:
    for c in custlist:
        if c[0] == d[2]:
            finalList.append(d)



customerDf = pd.DataFrame(finalList, columns=['CONTACTID','CUSTOMERID'
                                              ,'CUSTOMERNAME', 'PHONE'])

customerDf.to_csv("CUSTOMERDETAILS.csv", index=False)
print("CUSTOMERDETAILS.csv has been created")
```

Extracting customer details as a part of dimension table

Business Intelligence

Managers think of the business in terms of business dimensions.

Marketing Vice President

- How much did my new product generate?

month by month, in the southern division, by user demographic, relative to the previous version

Marketing Manager

- What are the sales statistics?

by products, summarized by product categories, daily, weekly, by distribution channels

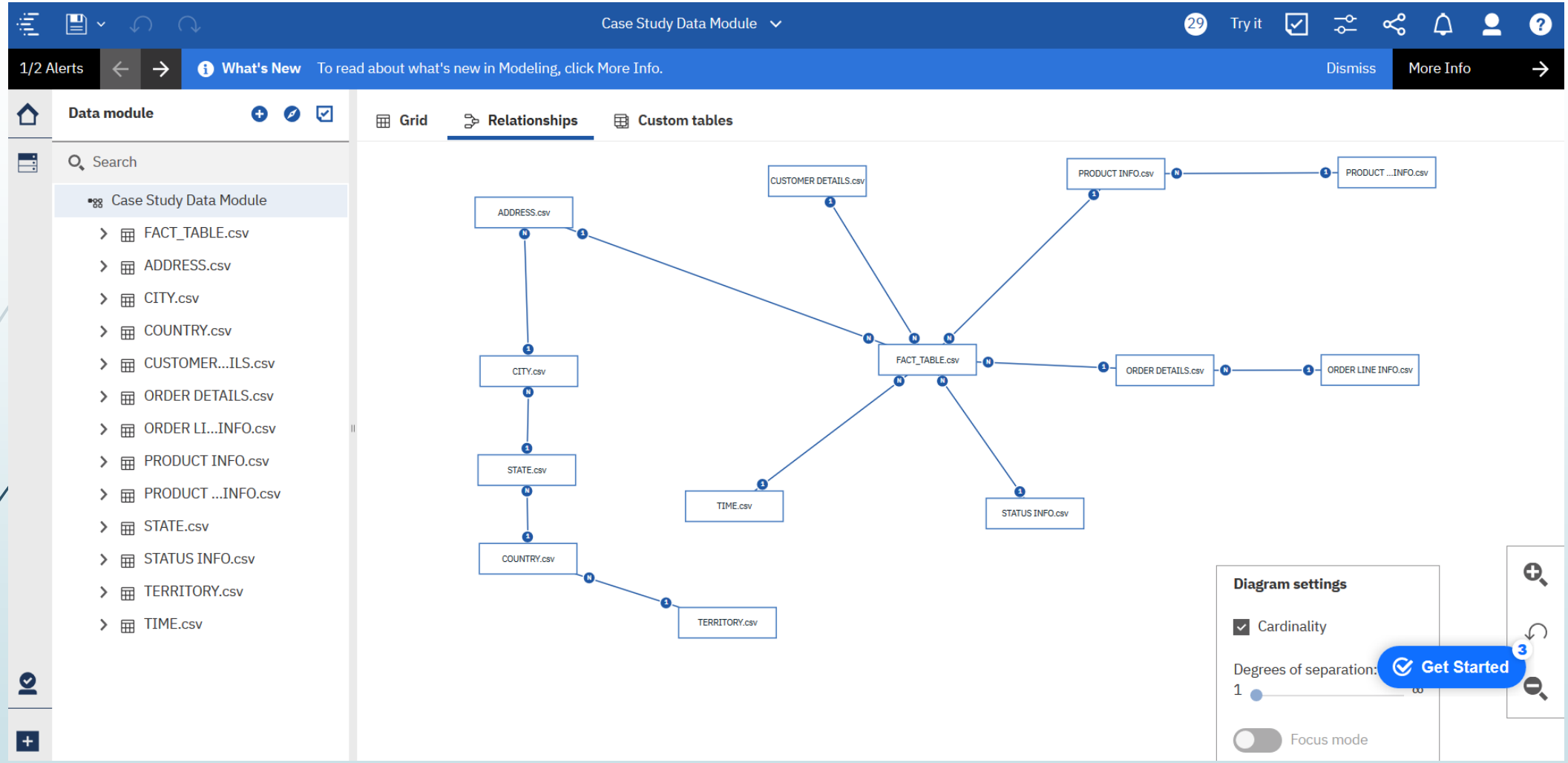
Financial Controller

- What are the expenses incurred?

Profit/loss

Business Intelligence

- It refers to technologies, applications and practices for presenting a business information to collect meaningful information.
- The tools used in BI will predict the views of business operations by analyzing the historical information.
- It helps to directly load data from the csv or cloud premises.
- IBM Cognos Analytics tool:
- Cognos Analytics is an AI-fueled business intelligence platform that supports the entire analytics cycle, from discovery to operationalization.
- Visualize, analyze and share actionable insights about your data with anyone in your organization.


Case Study Data Module
29
Try it
1/2 Alerts
What's New
To read about what's new in Modeling, click More Info.
Dismiss
More Info
Data module
Search
Case Study Data Module
FACT_TABLE.csv
ADDRESS.csv
CITY.csv
COUNTRY.csv
CUSTOMER...ILS.csv
ORDER DETAILS.csv
ORDER LI...INFO.csv
PRODUCT INFO.csv
PRODUCT ...INFO.csv
STATE.csv
STATUS INFO.csv
TERRITORY.csv
TIME.csv
Grid
Relationships
Custom tables
CUSTOMER DETAILS.csv
PRODUCT INFO.csv
PRODUCT ...INFO.csv
ADDRESS.csv
FACT_TABLE.csv
ORDER DETAILS.csv
ORDER LINE INFO.csv
CITY.csv
STATE.csv
TIME.csv
STATUS INFO.csv
COUNTRY.csv
TERRITORY.csv
Diagram settings
Cardinality
Degrees of separation:
1
Focus mode
Get Started
3

- We analyzed data and found insights, which we combined a single dashboard.
- There are some insights like:

- Product vs Sales data w.r.t Time and Status
- Sales vs Quarter w.r.t Time
- Sales for customer w.r.t Time
- Product vs Sales w.r.t Time and Client
- Sales vs Year w.r.t Status

References

- https://panoply.io/data-warehouse-guide/3-ways-to-build-an-etl-process/
- https://www.minerra.net/business-analytics/what-data-warehouse-how-deliver-value/
- https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm
- https://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/
- https://www.ibm.com/ca-en/products/cognos-analytics
- https://books.google.ca/books?id=n2nIM0l1TQ0C&pg=PA91&lpg=PA91&dq=Managers+think+of+the+business+in+terms+of+business+dimensions&source=bl&ots=sUR-HRNUfT&sig=ACfU3U3vaSp-gm-wu18ljwZ2YW9TGffdlg&hl=en&sa=X&ved=2ahUKEwjS99jSk8foAhXTYDUKHejhBlgQ6AEwCXoECA4QLg#v=onepage&q=Managers%20think%20of%20the%20business%20in%20terms%20of%20business%20dimensions&f=false