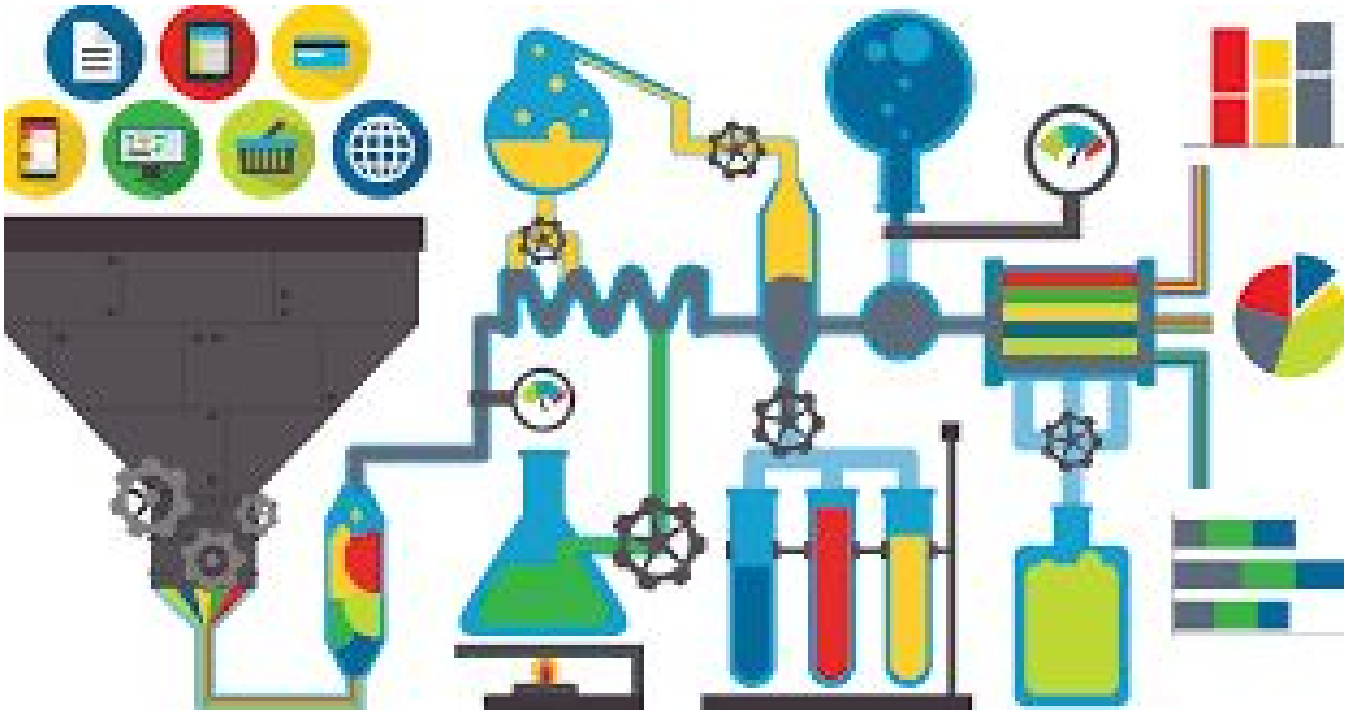
DATA WAREHOUSE
IST 722
Prof. P D Taber

WORLD MART

Online retail store management System



Shubham Shete
Harsh Darji
Harper He

World Mart- Company Description

World Mart is a Nationwide located retail corporation that operates on a chain of hypermarkets and grocery Stores. Every Day Low Price (EDLP) is the cornerstone of our strategy, and our price focus has never been stronger. Today's customer seeks the convenience of one-stop shopping that we offer. From grocery to all our customer and we provide the deep assortment that our customers appreciate.

From our humble beginnings as a small discount retailer in Syracuse, New York. World Mart has opened thousands of stores in the U.S. and expanded internationally. Through innovation, we're creating a seamless experience to let customers shop anytime and anywhere online, through mobile devices and in stores.

World Mart has stores in 50 states and Puerto Rico offering low prices on the broadest assortment of products through a variety of formats including the Supercenter, Discount Store and Neighborhood Market.

Data Warehouse Mission Statement

At World Mart, we research, identify and make available new products and categories that suit the everyday needs of the families. Our mission is to provide the best value possible for our customers, so that every dollar they spend on shopping with us gives them more value for money than they would get anywhere else.

World Mart is a one-stop supermarket chain that aims to offer customers a wide range of basic home and personal products under one roof. Each World Mart store stocks home utility products - including food, groceries and many more - available at competitive prices that our customers appreciate. Our core objective is to offer customers good products at great value.

Vision :

" To know what the customer wants and deliver it at the best price"

Mission Statement :

To be the lowest priced retailer in the area of operation and grocery. It is our continuous endeavor to investigate, identify and make available new product categories for customer's everyday use and at the best market value.

The objective behind building a data warehouse is to connect data from the retail systems, extracting and transforming it to store in one repository for improved data flow across the organization. This helps in better analytics considering vast data extract from different stores and result from discrete sources can be used to produce better business decisions in the retail market that would give a competitive edge over the business competitors.

Business Case

Every small and big business has used warehousing to improve its outreach, customer count, sales, profit and each possible aspect. But this was not sufficient. As data grew from megabytes to gigabytes to petabytes, these smart business felt a to store this data efficiently and to utilize it for improving various aspect of business. One such domain is retail where customer are products are key aspect. Which product is needed by what type of customer and when are the key questions of retail business. If they are answered well the can take retail business to new heights. In solving these queries Data Warehouse plays an important role. It helps to analyze key aspects to improve sale of retail stores. To know what customer buys and in which season, we need to have a look over the whole data. So first we need to collect the whole historical data in one place in a standard format. This is done by preparing data warehouse. There are many software which helps in this like Teradata, Netezza, Oracle, Hadoop, etc. Once the warehouse is prepared we can use this dataset in many ways to answer endless queries. In this project I have simulated the real time data warehouse preparation and answering business queries. Use of ETL is an integral part of the warehousing process for the extracting, transformation and loading process of the data.
Using the warehousing technique analytics on historical data can be done with ease that will help the decision support systems to provide better insights for the organizational growth. Ad-Hoc reporting and dashboard creation for the generation of attribute, location, product specific data can be done using warehousing techniques.

Project Overview

The project is focused towards creating a centralized data warehouse repository that consists of aggregated data coming from different stores. The data warehouse databases provide a decision support system in which user can calculate and evaluate the performance of the organization over time. The data will be stored in a series of snapshots, in which each record represents data at a specific time. By analyzing the snapshots, one can compare among the time periods. These comparisons can help user make important business decisions.
A Data warehouse stores the data organization wide like, customers, employees, products, locations, date and time. Data Mart are used to departmentalize these different attributes for analyzing the change in business based on their period data snapshots. The is denormalized to improve query performance. The collection from heterogenous sources shall be converted into one standard form for evaluation and analysis. This way the extracted data is transformed and then based on our warehousing approach it can be loaded into the warehouse.
The Integration and loading of data is a smooth functionality that provides ease of utilization. The main reason to implement warehouse is to provide better business decisions, departmentalization of data, product and customer analysis and increased organizational value.

Project Scope – InScope and Out of Scope

In scope :

- Building a warehouse that provides an estimate of the sales and CRM based on the historical data.
- A system that generates KPIs for World Mart based on the historic data that can be used for analytics
- Module that includes Datawarehouse, Data mart and Decision Support System Generation
- Integration Testing and Beta support testing Mechanism
- IT training to the employees and End User training pre-deployment phase
- A maintenance plan to provide support to the employees in cases of system crashes and periodic consultation routines to ensure the smooth working of the system.

Out of scope :

- Every Business Insight given by the DSS cannot assure an increased sales or customer base, however it can predict a near possibility to what sales might be done in the future.
- Erratic change in the behavior of the data due to increased/ decreased sale that are irregular or sporadic data might cause the DSS prediction to go wrong, thus the data inconsistency is not a factor that a warehouse can handle.

Estimated Cost and ROI

The cost breakdown for the project is mentioned below:

Major Costs:

1. Storage: 2000$ per month * 12 months = 24000$ annually
2. All Software: 1000$ per month * 12 months = 12000$ annually
3. Human Resources: 36,000$ per month (Team of 6) * 12 months = 432,000$

Additional Costs: 50,000$
Total Costs: 516,000$ approximately

Estimated ROI = Company currently has 10000 customers and has a growth of 5 percent with the implementation i.e. 500 and minimum order is of 1800$ = 500 *1800 =900,000$

ROI = 74.4%

Project Team

Core Team	Major Responsibilities
Project Manager	Define, Plan, Control and review all project activities
Business Analyst	Gather Business Requirements, make business decisions, resolve disputes between business units and improve the source data quality
Design Specialist	Convert Business Requirements into system design. Establish and maintain the technical infrastructure(hardware, network,middleware, system software)
Data Warehouse Architect	Build, enhance,load and manipulate the metadata repository. Perform cross-organizational data analysis, establish naming standards, create the project specific logical data models and merge those models into an enterprise logical data model. Design and develop the ETL process
BI Architect	Connect the data warehouse to BI systems. Train the stakeholders to use the BI tools for reporting purposes.
SQL DBA	Design,load, monitor, and tune the data warehouse databases

Project Stakeholders

Our stakeholders in this project can be broadly categorized in the following four buckets:

1. Project Sponsor: Director of World Mart company who is funding the project
2. Project Team: Team responsible for implementing the data warehouse
3. End Users: Employees using the data warehouse for data analytical purposes
4. Review Board: The team responsible to check whether the DW meets the requirements

Interview Questions:

Questions	Stakeholder(s) to be asked
What are the data sources and type of data?	**Data Administrator**
What regulations do we need to adhere to?	**Legal Team at World Mart**
What kind of reporting and analytical services it should provide?	**Business Analyst at World Mart, End Users**
Do we need Data marts or just one consolidated DW?	**Project Team, End Users**
What are the existing data security measures?	**Data Administrator, Chief Technical Officer**
Do you need the team to support the data warehouse or it can be done in house?	**Chief Technical Officer**
What are current problems while extracting data from different sources for analytical purposes?	**Data Analyst**

Design

Bus Matrix

Business Process Name	Fact Table	Fact Grain Type	Granularity	Facts	Customer	Employee	Product	Date
Sales	Fact_Sales	Transaction	One row per sales detail	Quantity, Cost, Amount, Total Amount	X	X	X	X

Data Dictionary

Below is the list of all attributes we used in our data warehouse.

Table	Columns	Contents	Data Type	Required?	PK/FK
Dim_Product	Pro_Id	Primary Key	int	Y	PK
	Pro_Code	Business Key	int	Y	
	Pro_Type	Type of product	nvarchar	Y	
	Pro_Name	Name of product	nvarchar	Y	
	Pro_Unit	Unit of product	nvarchar	Y	
Dim_Employee	Emp_Id	Primary Key	int	Y	PK
	Emp_SSN	Business Key	varchar	Y	
	Emp_Last_Name	Last name of the employee	nvarchar	Y	
	Emp_First_Name	First name of the employee	nvarchar	Y	
	Emp_Birthdate	Date of birth of the employee	datetime	Y	
	Emp_Gender	Gender of the employee	nvarchar	Y	
	Emp_Email	Email of the employee	nvarchar	Y	
	Emp_Phone	Phone number of the employee	nvarchar	Y	
	Emp_Zip	Zipcode of the employee	nvarchar	Y	
	Emp_Address	Address of the employee	nvarchar	Y	
	Emp_City	City of the employee	nvarchar	Y	
	Emp_State	State of the employee	nvarchar	Y	
	Emp_Salary	Salary of the employee	int	Y	
	Emp_Position	Position of the employee	nvarchar	Y	

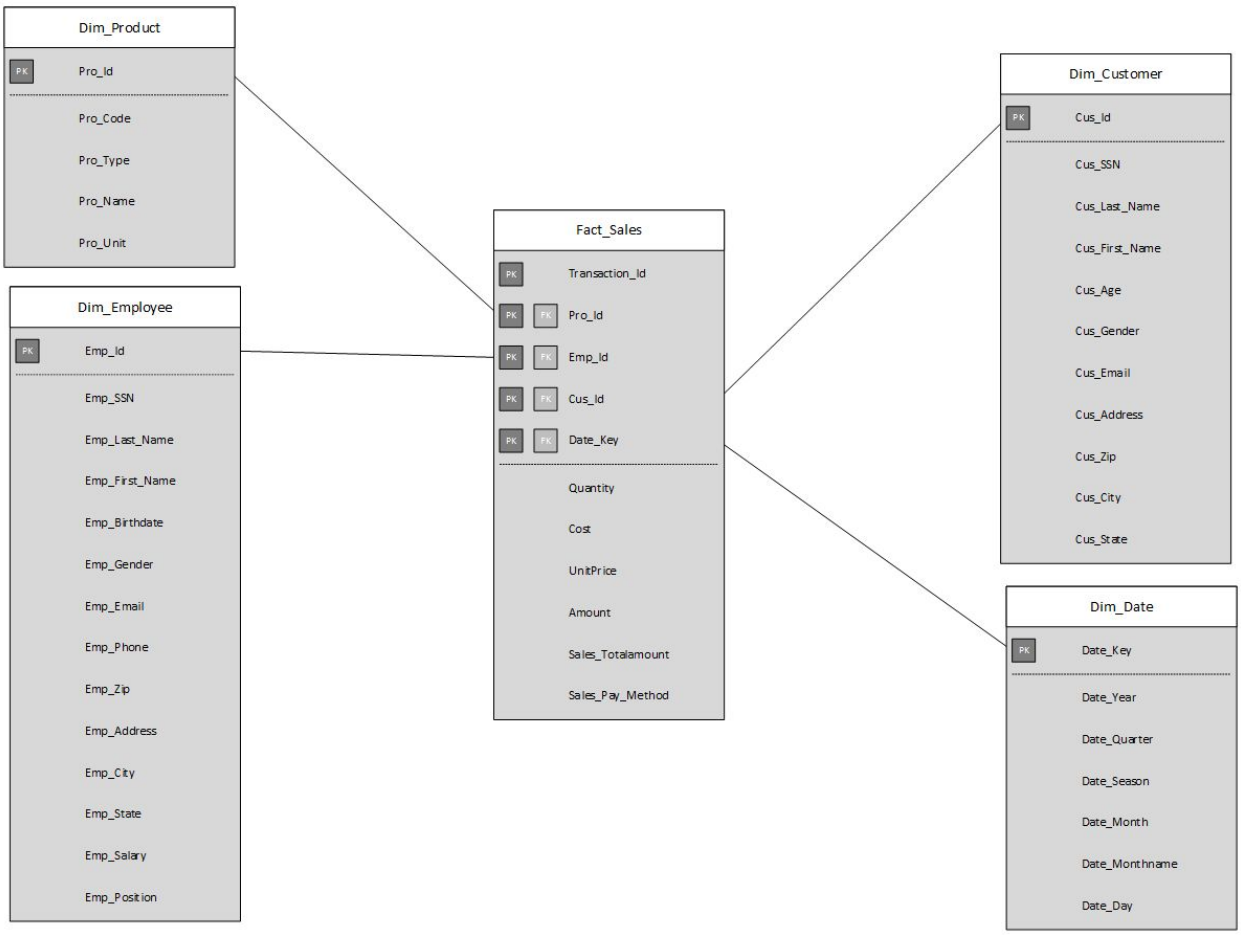
Table	Columns	Contents	Data Type	Required?	PK/FK
Dim_Customer	Cus_Id	Primary Key	int	Y	PK
	Cus_SSN	Business Key	varchar	Y	
	Cus_Last_Name	Last name of the Customer	nvarchar	Y	
	Cus_First_Name	First name of the Customer	nvarchar	Y	
	Cus_Age	Age of the Customer	int	Y	
	Cus_Gender	Gender of the Customer	nvarchar	Y	
	Cus_Email	Email of the Customer	nvarchar	Y	
	Cus_Phone	Phone number of the Customer	nvarchar	Y	
	Cus_Address	Address of the Customer	nvarchar	Y	
	Cus_Zip	Zipcode of the Customer	nvarchar	Y	
	Cus_City	City of the Customer	nvarchar	Y	
	Cus_State	State of the Customer	nvarchar	Y	
Dim_Date	Date_Key	Primary Key	int	Y	PK
	Date_Year	Year	int	Y	
	Date_Quarter	Quarter	varchar	Y	
	Date_Season	Season	varchar	Y	
	Date_Month	Month	int	Y	
	Date_Month_Name	Month Name	varchar	Y	
	Date_Day	Day	int	Y	
Fact_Sales	Transaction_Id	Primary Key	int	Y	PK
	Pro_Id	Primary Key	int	Y	PK/FK
	Emp_Id	Primary Key	int	Y	PK/FK
	Cus_Id	Primary Key	int	Y	PK/FK
	Date_Key	Primary Key	int	Y	PK/FK
	Quantity	Quantity of this product in this order	int	Y	
	Cost	Cost of this product in this order	decimal	Y	
	UnitPrice	Unit Price of this product in this order	decimal	Y	
	Amount	Amount of this product in this order	decimal	Y	
	Sales_Total_Amount	Total amount of this order	decimal	Y	
	Sales_Pay_Method	Payment method of this order	varchar	Y	

Issue List

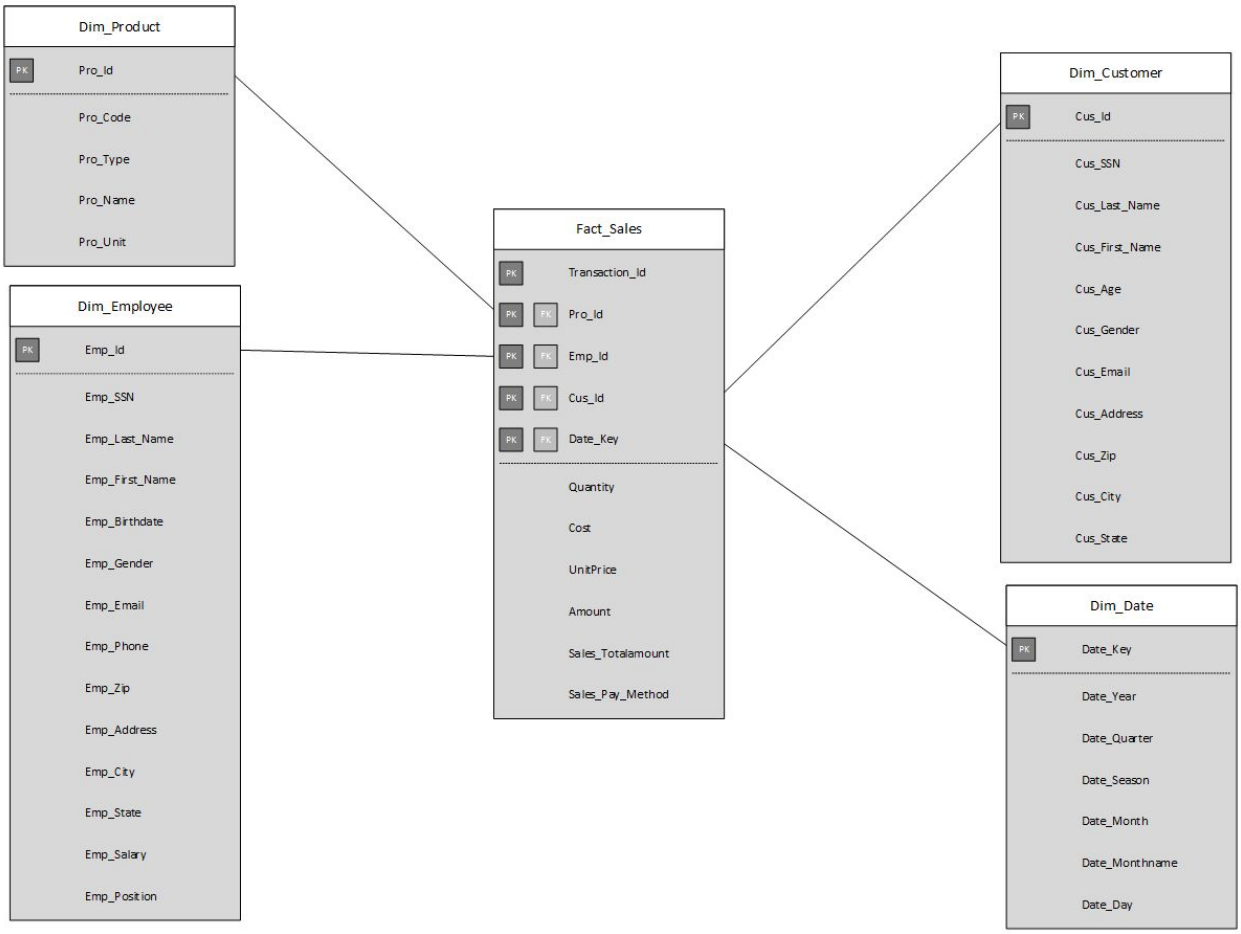
Issue #	Task/Topic	Issue	Identified Date	Reported By	Responsible	Status	Priority
1	Data Security	The data contains SSN so it is important to keep our data secured.	Thu 4/25/19	Team Lead	Team Lead	Closed	High
2	Performance	Is the project worth the investment or not?	Thu 4/25/19	Project Sponsor	Team Lead	Closed	High
3	User Acceptance	The system should be user firendly.	Tue 5/7/19	Team Lead	Team Lead	Closed	High
4	Design	The initial design should be optimal else going back and forth iscrease pproject costs.	Tue 5/7/19	Team Member	Team Member	Closed	High
5	Evaluation	How to evaluate the project?	Thu 4/25/19	Project Sponsor	Team Lead	Closed	Medium
6	Maintenance	How to maintain the project?	Fri 4/19/19	Team Member	Team Member	Closed	Medium

Dimensional Models and Schemas

Star Schema of Sales information



Data Warehouse Schemas Design



Source to Target Analysis

- Identified all source systems that have data to be extracted from
- Identified valuable data and formed aggregation/summarization strategies
- Identified data types and their corresponding data types in the ETL tool used for the process
- Summarized data in such a way that valuable information is not lost, and storage costs are kept minimum

SQL Scripts

1. Dim_Product

```
USE [ist722_xhe128_dw]
GO
/****** Object: Table [dbo].[Dim_Product] Script Date: 04/28/2019
11:16:28 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[Dim_Product](
[Pro_Id] [INT] NOT NULL,
[Pro_Code] [INT] NULL,
[Pro_Type] [NVARCHAR] (50) NULL,
[Pro_Name] [NVARCHAR] (50) NULL,
[Pro_Unit] [NVARCHAR] (50) NULL,
CONSTRAINT [PK_Dim_Product] PRIMARY KEY CLUSTERED
(
[Pro_Id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

2. Dim_Employee

```
USE [ist722_xhe128_dw]
GO
/****** Object: Table [dbo].[Dim_Employee] Script Date: 04/28/2019
11:17:43 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE[dbo].[Dim_Employee](
[Emp_Id] [INT] NOT NULL,
[Emp_SSN] [nvarchar](50) NULL,
[Emp_Last_Name] [nvarchar](50) NULL,
[Emp_First_Name] [nVARCHAR] (50) NULL,
[Emp_Birthdate] [datetime] NULL,
[Emp_Gender][NVARCHAR] (50) NULL,
[Emp_Email] [NVARCHAR] (50) NULL,
[Emp_Phone] [nvarchar](50) NULL,
```

```
[Emp_Address] [NVARCHAR] (50) NULL,
[Emp_Zip] [int] NULL,
[Emp_City] [NVARCHAR] (50) NULL,
[Emp_State] [VARCHAR] (50) NULL,
[Emp_Salary] [INT] NULL,
[Emp_Position] [VARCHAR] (50) NULL,
CONSTRAINT [PK_Dim_Employee] PRIMARY KEY CLUSTERED
(
[Emp_Id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

3. Dim_Customer

```
 USE [ist722_xhe128_dw]
GO
/****** Object: Table [dbo].[Dim_Customer] Script Date: 04/28/2019
11:27:24 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE[dbo].[Dim_Customer](
[Cus_Id] [INT]  NOT NULL,
[Cus_SSN]  [nvarchar](50) NULL,
[Cus_Last_Name] [nvarchar](50) NULL,
[Cus_First_Name] [nVARCHAR] (50) NULL,
[Cus_Age] [INT] NULL,
[Cus_Gender] [NVARCHAR] (50) NULL,
[Cus_Email] [NVARCHAR] (50) NULL,
[Cus_Phone] [nvarchar](50) NULL,
[Cus_Address] [NVARCHAR] (50) NULL,
[Cus_Zip] [int] NULL,
[Cus_City] [NVARCHAR] (50) NULL,
[Cus_State] [VARCHAR] (50) NULL,
CONSTRAINT [PK_Dim_Customer] PRIMARY KEY CLUSTERED
(
[Cus_Id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

USE [ist722_xhe128_dw]
GO
/****** Object: Table [dbo].[Dim_Date] Script Date: 04/28/2019
11:40:34 AM ******/
```

```
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
```

4. Dim_Date

```
 CREATE TABLE [dbo].[Dim_Date](
[Date_Key] [INT]  NOT NULL,
[Date_Year] [INT]  NULL,
[Date_Quarter] [VARCHAR](2) NULL,
[Date_Season] [VARCHAR](6) NULL,
[Date_Month] [INTEGER]  NULL,
[Date_Monthname] [VARCHAR](3) NULL,
[Date_Day] [INTEGER]  NULL,
CONSTRAINT [PK_Dim_Date] PRIMARY KEY CLUSTERED
(
[Date_Key] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

5. Fact_Sales

```
USE [ist722_xhe128_dw]
GO
/****** Object: Table [dbo].[Fact_Sales] Script Date: 04/28/2019
12:45:53 AM ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE[dbo].[Fact_Sales](
[Transaction_Id] [INT] NOT NULL,
[Pro_Id] [INT] NULL ,
[Emp_Id] [INT] NULL ,
[Cus_Id] [INT] NULL ,
[Date_Key] [INT] NULL ,
[Quantity] [INT] NULL ,
[Cost] [decimal](10,2) NULL ,
[UnitPrice] [decimal](10,2) NULL ,
[Amount] [decimal](10,2) NULL ,
[Sales_Totalamount] [decimal](10,2) NULL ,
[Sales_Pay_Method] [VARCHAR](100) NULL,

CONSTRAINT [PK_Fact_Sales] PRIMARY KEY CLUSTERED
(
[Transaction_Id]ASC
```

```
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY
= OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
ALTER TABLE [dbo].[Fact_Sales] WITH CHECK ADD CONSTRAINT
[FK_Fact_Sales_C] FOREIGN
KEY([Cus_Id])
REFERENCES [dbo].[Dim_Customer] ([Cus_Id])
GO

ALTER TABLE [dbo].[Fact_Sales] CHECK CONSTRAINT
[FK_Fact_Sales_C]
GO
ALTER TABLE [dbo].[Fact_Sales] WITH CHECK ADD CONSTRAINT
[FK_Fact_Sales_P] FOREIGN
KEY([Pro_Id])
REFERENCES [dbo].[Dim_Product] ([Pro_Id])
GO

ALTER TABLE [dbo].[Fact_Sales] CHECK CONSTRAINT
[FK_Fact_Sales_P]
GO
ALTER TABLE [dbo].[Fact_Sales] WITH CHECK ADD CONSTRAINT
[FK_Fact_Sales_D] FOREIGN
KEY([Date_Key])
REFERENCES [dbo].[Dim_Date] ([Date_Key])
GO

ALTER TABLE [dbo].[Fact_Sales] CHECK CONSTRAINT
[FK_Fact_Sales_D]
GO
ALTER TABLE [dbo].[Fact_Sales] WITH CHECK ADD CONSTRAINT
[FK_Fact_Sales_E] FOREIGN
KEY([Emp_Id])
REFERENCES [dbo].[Dim_Employee] ([Emp_Id])
GO

ALTER TABLE [dbo].[Fact_Sales] CHECK CONSTRAINT
[FK_Fact_Sales_E]
GO
```

SQL Insert

Dim_Product

	Pro_Id	Pro_Code	Pro_Type	Pro_Name	Pro_Unit
1	1	4011	Bread	Breadcrumbs	1 kg
2	2	4012	Bread	Panini	each
3	3	4013	Dairy	Butter	1.8 kg
4	4	4014	Dairy	Buttermilk	1 litre
5	5	4015	Dairy	Cream	0.5 litre
6	6	4016	Dairy	Milk	1 litre
7	7	4017	Fruit	Apple	each
8	8	4018	Fruit	Avocado	each
9	9	4019	Fruit	Banana	1 each
10	10	4020	Fruit	Blackberries	1 kg

Dim_Employee

	Emp_Id	Emp_SSN	Emp_Last_Name	Emp_First_Name	Emp_Birthdate	Emp_Gender	Emp_Email	Emp_Phone	Emp_Address	Emp_Zip	Emp_City	Emp_State	Emp_Salary	Emp_Position
1	1	788-288-718	lawson	chris	1988-02-22 00:00:00.000	f	Chri.LAWSON7987@mail2web.com	(701) 446-2519	8664 haddock	35027	birmingham	al	3986	Clerk
2	2	194-651-816	solis	isaac	1968-08-08 00:00:00.000	f	Is.SOLIS4398@yahoo.com	(539) 946-9174	5308 pratt	10633	new york	ny	4626	Cashier
3	3	684-754-990	joyner	delilah	1972-09-16 00:00:00.000	m	Deli.JOY5182@gmail.com	(763) 937-3965	9663 ob dan ryan	80277	lakewood	co	5278	Supervisor
4	4	328-874-858	hampton	hana	1981-02-03 00:00:00.000	m	Ha.HAMPTO4382@mail2web.com	(304) 706-7545	2969 deming	31313	savannah	ga	4658	Cashier
5	5	363-130-194	flynn	abbigail	1994-07-04 00:00:00.000	m	Abbig.FLY6680@mail2web.com	(972) 447-4596	2985 ob dan ryan	48871	lansing	mi	4659	Cashier
6	6	377-542-826	oneal	zaiden	1976-05-18 00:00:00.000	m	Zaiden.ON1961@mail2web.com	(812) 508-4906	4982 luella	95305	modesto	ca	5858	Manager

Dim_Customer

	Cus_Id	Cus_SSN	Cus_Last_Name	Cus_First_Name	Cus_Age	Cus_Gender	Cus_Email	Cus_Phone	Cus_Address	Cus_Zip	Cus_City	Cus_State
1	1	640-739-331	sweeney	abdiel	61	m	Abdi.SWE3222@gmail.com	(440) 678-0296	6515 avenue k	48312	sterling heights	michigan
2	2	967-415-725	delgado	asia	28	m	As.DEL3467@mail2web.com	(224) 793-6141	323 ogallah	92818	anaheim	california
3	3	123-788-538	chapman	kayla	60	f	Ka.CHAPMAN6294@hushmail.com	(609) 106-7048	4502 artesian	15651	pittsburgh	pennsylvania
4	4	553-991-446	mack	kallie	40	m	Kall.MA7941@gmail.com	(603) 621-9708	1741 london	72748	fayetteville	north carolina
5	5	961-515-413	levine	mara	42	m	Ma.LEV2307@live.com	(484) 232-1424	5887 hillock	29402	charleston	south carolina
6	6	085-224-132	reilly	keyla	35	f	Keyl.REILLY4085@hushmail.com	(775) 180-1724	382 lad sb stevenson ob	37553	knoxville	tennessee
7	7	706-956-751	mathis	august	63	m	August.MATHIS7741@live.com	(657) 533-2858	6066 kewanee	50649	montpelier	vermont
8	8	675-714-302	tyson	jacqueline	25	m	Jacqueline.TYSON7863@yahoo.com	(304) 643-1086	9103 columbus lower	37489	chattanooga	tennessee
9	9	911-144-729	figueroa	ariel	33	f	Ariel.FIGUERO1233@mail2web.com	(319) 694-5097	3642 52nd	79338	el paso	texas
10	10	586-185-435	sanders	samiyah	72	f	Sami.SANDERS5686@gmail.com	(626) 855-7603	6972 drew	71123	shreveport	louisiana

Dim_Date

	Date_Key	Date_Year	Date_Quarter	Date_Season	Date_Month	Date_Monthname	Date_Day
1	1	2018	Q1	winter	1	Jan	1
2	2	2018	Q1	winter	2	Feb	8
3	3	2018	Q1	spring	3	Mar	13
4	4	2018	Q1	spring	3	Mar	28
5	5	2018	Q2	spring	4	Apr	6
6	6	2018	Q2	spring	4	Apr	8
7	7	2018	Q2	spring	4	Apr	10
8	8	2018	Q2	spring	5	May	11
9	9	2018	Q2	spring	5	May	15
10	10	2018	Q2	summer	6	Jun	15

Fact_Sales

	Transaction_Id	Pro_Id	Emp_Id	Cus_Id	Date_Key	Quantity	Cost	UnitPrice	Amount	Sales_Totalamount	Sales_Pay_Method
1	901	2	3	2	3	2	0.30	0.37	0.74	7.94	Credit Card
2	902	3	3	2	3	1	5.76	7.20	7.20	7.94	Credit Card
3	903	4	4	2	5	3	0.62	0.78	2.34	2.34	Cash
4	904	1	1	1	1	3	1.92	2.40	7.20	7.20	Credit Card
5	905	5	1	1	1	4	0.71	0.89	3.56	3.56	Cash
6	906	6	2	4	2	2	0.47	0.59	1.18	1.18	Debit Card
7	907	7	5	5	6	3	0.28	0.35	1.05	1.05	Cash
8	908	8	6	6	7	4	0.68	0.85	3.40	3.40	Debit Card
9	909	9	6	7	4	10	0.23	0.29	2.90	2.90	Cash
10	910	10	4	5	9	2	3.04	3.80	7.60	7.60	Credit Card
11	911	10	5	9	8	3	3.04	3.80	11.40	11.40	Credit Card

SSIS Transformations

1. **Fuzzy Grouping**

In the customer table, some of the state named have been inserted wrongly. So, we use fuzzy grouping transformation to correct the state name. Eg: new.York, new.york is changed to new york. The output can be found in the following images.

Before Transformation:

	Customer_Pho...	Customer_Address	Customer_...	Customer_City	Customer_State
m	(609) 106-7048	4502 artesian	15651	pittsburgh	new.YORK
	(224) 793-6141	323 ogallah	92818	anaheim	new.york
	(440) 678-0296	6515 avenue k	48312	sterling heights	newyork
	(630) 619-9282	8840 grand	84050	provo	utah
	(479) 714-4073	8769 burnham	48146	detroit	michigan
	(772) 356-5717	3719 65th	58033	fargo	north dakota
n	(810) 737-1947	2719 landers	60488	joliet	illinois
	(772) 210-4557	1427 keokuk	46374	indianapolis	indiana
	(707) 146-9937	7895 berteau	29408	charleston	south carolina
ı	(801) 964-5258	5428 mannheim	14686	rochester	new york

Data Flow:

After Transformation:

Customer_...	Customer_City	Customer_St...	Customer_State_clean	_Similarity_Customer_S
15651	pittsburgh	new.YORK	new york	0.9875
92818	anaheim	new.york	new york	0.9875

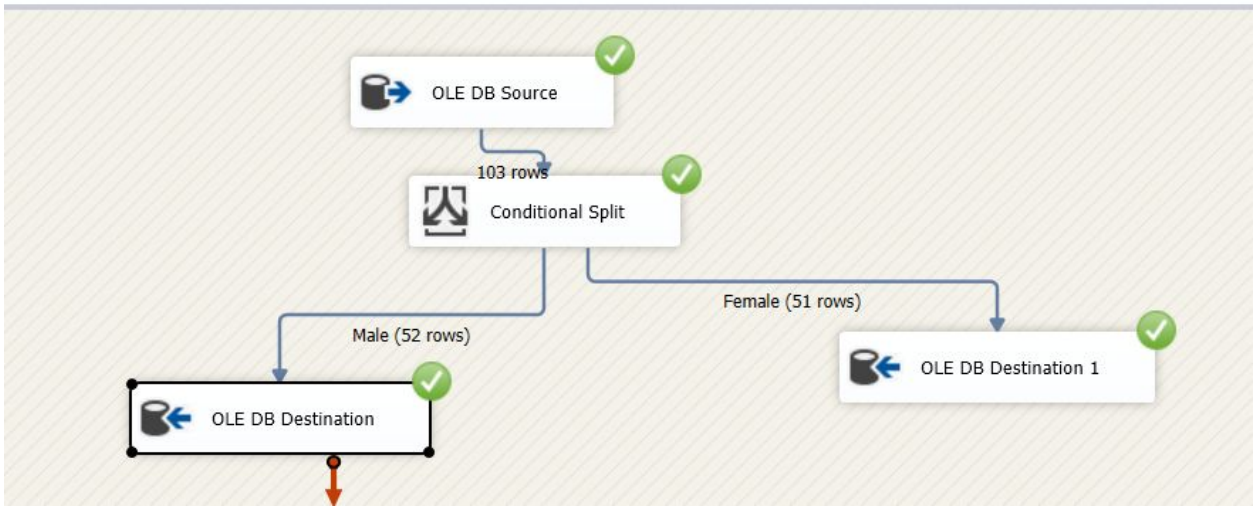
2. CONDITIONAL SPLIT

We often want to analyze our customers depending on their gender. So, here we haved used conditional split to separate male and female customers which will help us to analyze our customers better. The output can be found in images below.

BEFORE:

Customer_Lastna...	Customer_Firstna...	Customer_A...	Customer_Gen...	Customer_Email	C
sweeney	abdiel	61	m	Abdi.SWE3222@gmail.com	(
delgado	asia	98	m	As.DEL3467@mail2web.com	(
chapman	kayla	60	f	Ka.CHAPMAN6294@hushmail.com	(
mack	kallie	80	m	Kalli.MA7941@gmail.com	(
levine	mara	42	m	Ma.LEV2307@live.com	(
reilly	keyla	75	f	Keyl.REILLY4085@hushmail.com	(
mathis	august	63	m	August.MATHIS7741@live.com	(
tyson	jacqueline	55	m	Jacqueline.TYSON7863@yahoo.com	(
figueroa	ariel	33	f	Ariel.FIGUERO1233@mail2web.com	(
sanders	samiyah	72	f	Sami.SANDERS5686@gmail.com	(

DATA FLOW:



After:
Male Data

...r...	Customer_Lastna...	Customer_Firstna...	Customer_A...	Customer_Gen...	Customer_Email
	sweeney	abdiel	61	m	Abdi.SWE3222@gmail.com
	delgado	asia	98	m	As.DEL3467@mail2web.com
	mack	kallie	80	m	Kalli.MA7941@gmail.com
	levine	mara	42	m	Ma.LEV2307@live.com
	mathis	august	63	m	August.MATHIS7741@live.com
	tyson	jacqueline	55	m	Jacqueline.TYSON7863@yahoo.com
	spence	zariyah	12	m	Zariyah.SPENCE7681@mail2web.com
	wall	aviana	54	m	Avi.WALL7748@yahoo.com
	delacruz	kason	67	m	Kas.DELA3771@gmail.com
	gardner	aliya	49	m	Aliya.GAR4835@mail2web.com

Female Data:

	Customer...	Customer_Lastna...	Customer_Firstna...	Customer_A...	Customer_Gen...	Customer_Email
1	3	chapman	kayla	60	f	Ka.CHAPMAN6294@hushmail.c
2	6	reilly	keyla	75	f	Keyl.REILLY4085@hushmail.co
3	9	figueroa	ariel	33	f	Ariel.FIGUERO1233@mail2web
4	10	sanders	samiyah	72	f	Sami.SANDERS5686@gmail.co
5	11	logan	darian	89	f	Daria.LOGA5610@live.com
6	12	pennington	malachi	61	f	Mala.PENNIN7691@gmail.com
7	14	romero	deborah	26	f	Debo.ROMER8266@hushmail.c
8	20	burch	ramon	93	f	Ramon.BUR9871@live.com
9	21	wheeler	sawyer	86	f	Sawyer.WHEELER4174@gmail
10	27	gaines	bowen	36	f	Bo.GAINE5005@mail2web.com

3. Derived Columns:

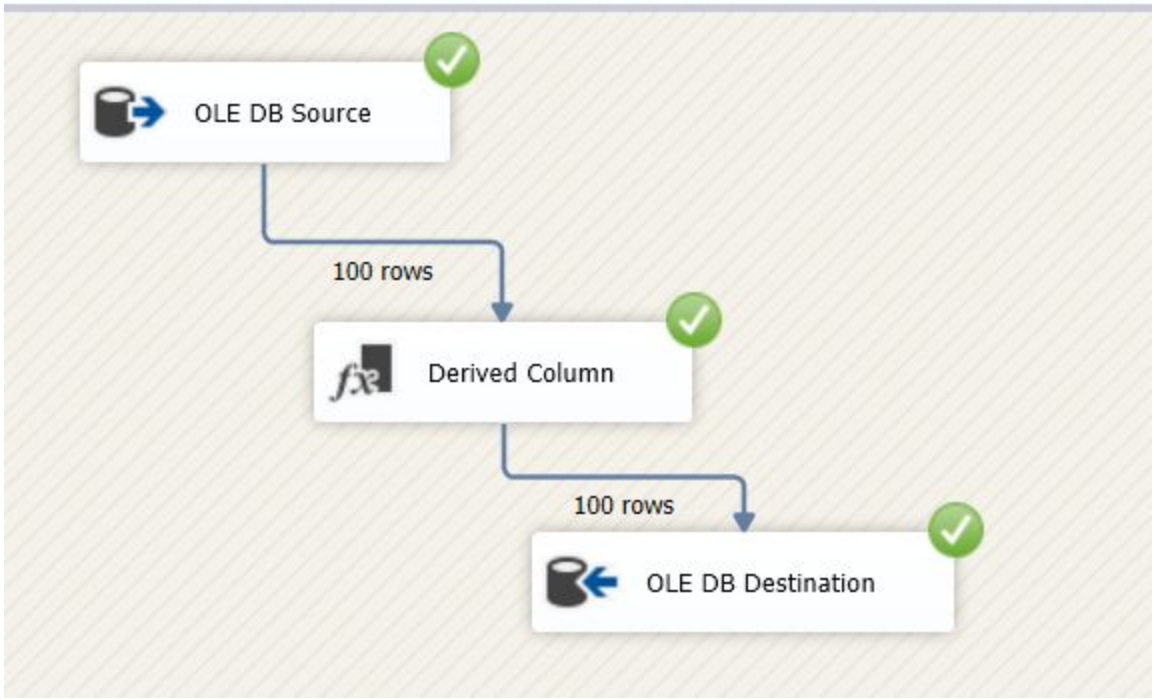
In our employee table there are five categories of employees- Clerk, cashier, customer representative, supervisor and manager. Here, manager and supervisor are Top Management and rest belong to general employees. So, we use derived column transformation to make a

new column named Hierarchy which distinguishes our employees. The output can be found in the mages below.

Before Transformation:

Employee_Email	Employee_Pho...	Employee_Address	Employee_...	Employee_City	Empl...	Employ...	Employee_Position
Chri.LAWSON7987@mail2web.com	(701) 446-2519	8664 haddock	35027	birmingham	al	29686	Clerk
Is.SOLIS4398@yahoo.com	(539) 946-9174	5308 pratt	10633	new york	ny	646268	Cashier
Deli.JOY5182@gmail.com	(763) 937-3965	9663 ob dan ryan	80277	lakewood	co	959278	Supervisor
Ha.HAMPTO4382@mail2web.com	(304) 706-7545	2969 deming	31313	savannah	ga	695658	Cashier
Abbig.FLY6680@mail2web.com	(972) 447-4596	2985 ob dan ryan	48871	lansing	mi	669659	Cashier
Zaiden.ON1961@mail2web.com	(812) 508-4906	4982 luella	95305	modesto	ca	555858	Manager
Kylie.CLI8243@yahoo.com	(682) 583-3085	5503 dan ryan 71st st	49287	jackson	mi	567830	Cashier
Pati.ODON8624@yahoo.com	(651) 828-0634	8340 post	40551	louisville	ky	341316	Supervisor
Princes.NAS7708@hushmail.com	(331) 751-8598	5761 mclean	49263	jackson	mi	626428	Clerk
Eliana.NELSO9646@yahoo.com	(507) 513-1119	4422 dayton	57051	sioux falls	sd	230320	Supervisor

Data Flow Diagram:



After Transformation:

	Employee_...	Employee_City	Employee_St...	Employee_Sal...	Employee_Positi...	Hierarchy
	35027	birmingham	al	29686	Clerk	General Employee
	10633	new york	ny	646268	Cashier	General Employee
	80277	lakewood	co	959278	Supervisor	Top Management
	31313	savannah	ga	695658	Cashier	General Employee
	48871	lansing	mi	669659	Cashier	General Employee
	95305	modesto	ca	555858	Manager	Top Management
it	49287	jackson	mi	567830	Cashier	General Employee
	40551	louisville	ky	341316	Supervisor	Top Management
	49263	jackson	mi	626428	Clerk	General Employee
	57051	sioux falls	sd	230320	Supervisor	Top Management

Maintenance Plan

Maintenance plan		
What	**Who**	**When**
Back up all system and user databases.	Administrative DBA	Every week on Monday
Update the record of all service packs that we install for both Microsoft Windows NT Server and Microsoft SQL Server. Keep records of the network libraries, the security mode, and the system administrator password.	Administrative DBA	Every week on Monday
Practice system and data recovery steps ahead of time on another server, and modify the steps as necessary to adapt to the environment.	Data Architect	The first Monday of each month
Create a data exposure analysis that defines downtimes for recovery and any potential data loss from each possible system failure.	Development DBA	Every three months on the first Tuesdayof the month
Reducing the size of data files by removing empty database pages. Reducing file size utilizes disk space more efficiently.	Administrative DBA	Every three months on the first Tuesdayof the month
Updating index statistics to ensure that the query optimizer has current information regarding the distribution of data values in the tables.	Development DBA	Every three months on the first Tuesdayof the month
Performing internal consistency checks of the data and data pages within the database to ensure that a system or software problem has not damaged data.	Data Architect	The first Monday of each month

Evaluation Plan

Evaluation plan					
What	**Meaning**	**Unit**	**Standard**	**Who**	**When**
Computing budget (CB)	Sum of operating budgets for software, hardware, and communication to support data warehouse operations	Monthly budget ($)	3000	Project Manager	The first Monday of each month
Labor budget (LB)	Labor to support data warehouse operations	Monthly direct budget ($)	36000	Project Manager	The first Monday of each month
Availability (Av)	Hours of service for user queries; A weighted measure should be used if parts of the warehouse have different availabilities.	Hours per day	14 hours (8am-10pm)	Project Manager	Every week on Monday
Queries (NQ)	Number of data requests either directly through ad hoc queries or indirectly through execution of planned reports.	Number of queries per month	1000	Database Administrator	The first Tuesday of each month
Flexibility ratio (FR)	Indicates the relative number of ad hoc queries to scheduled queries	Ratio of unplanned to planned queries per month	50%	Project Manager	The first Tuesday of each month
Number of users (NU)	Users who login to a data warehouse site at least once per month	Number of active users per month	20	Database Administrator	Every three months on the first Tuesdayof the month
Data age (DA)	Indicates the daily refresh interval for the data warehouse.	Weighted daily refresh interval in hours	1 hours	Database Administrator	The first Monday of each month

Conclusion

This newly formed central repository i.e. data warehouse is summarized data coming from various sources, without loss of any valuable information which is used for better decision making across the organization.