



Project Report

IMPLEMENTING SEVERAL TECHNIQUES FOR DATA ANALYTICS AND VISUALISATION

MODULE TITLE: DATA ANALYTICS AND VISUALISATION

MODULE CODE: B9IS107

GROUP MEMBERS:

SINGUPURAM NAVIN KUMAR | 10612366

MODULE LEADER: DR SHAZIA A AFZAL

TABLE OF CONTENTS

1. INTRODUCTION.....	2
1.1. REASONS FOR SELECTING THE SUBJECT AREA AND DATA.....	4
1.2. VISION AND GOALS	4
1.3 KEY STAKEHOLDERS	5
1.4 BUSINESS REQUIREMENTS.....	5
2. SCHEMA / DIMENSION MODEL	5
3. IMPLEMENTATION OF DATA WAREHOUSE	7
3.1 ETL TO POPULATE DATA WAREHOUSE	8
4. VISUALIZATIONS AND REPORTS.....	11
4.1 VISUALIZATIONS	12
4.2 REPORTS	16
5. CONCLUSION	21
6. References:	21
7. Appendix.....	22
SQL Queries	22
Report SP's	23



TABLE OF FIGURES

Figure 1: Database Schema	3
Figure 2: Star Schema for Employee Dataset	6
Figure 3: Dimension Model for the Dataset	7
Figure 4: Employee Dimension.....	8
Figure 5: Department Dimension	9
Figure 6: Calendar Dimension	10
Figure 7: Fact Salary Dimension	11
Figure 8: Database connection in Tableau	12
Figure 9: Average Salary in Years	13
Figure 10: Number of employees per job role.....	13
Figure 11: Average Salary per Role.....	14
Figure 12: Standard deviation of salary per role	15
Figure 13: Dashboard.....	16
Figure 14 : Department Gender Salary.....	17
Figure 15: Department Employee Count.....	18
Figure 16: Department Wise Salary.....	19
Figure 17: Department Title Wise Salary.....	20



1. INTRODUCTION

The Employee database is the dataset that was selected for the Data Analytics project. It is available on the [Relational Dataset Repository](#) website. Here are some specifics about the database:

- There are 6 tables in total:

1. Salaries

Columns are:

- (a) emp_no
- (b) salary
- (c) from_date
- (d) to_date

2. Titles

Columns are:

- (a) emp_no
- (b) title
- (c) from_date
- (d) to_date

3. Dept_emp

Columns are:

- (a) emp_no
- (b) dept_no
- (c) from_date
- (d) to_date

4. Dept_manager

Columns are:

- (a) dept_no
- (b) emp_no
- (c) from_date
- (d) to_date

5. Employees

Columns are:

- (a) emp_no
- (b) birth_date
- (c) first_name
- (d) last_name



- (e) gender
- (f) hire_date

6. Departments

Columns are:

- (a) dept_no
- (b) dept_name

- A total of 3,911,392 rows.
- A total of 24 columns.

This is the database schema, taken from the original website:

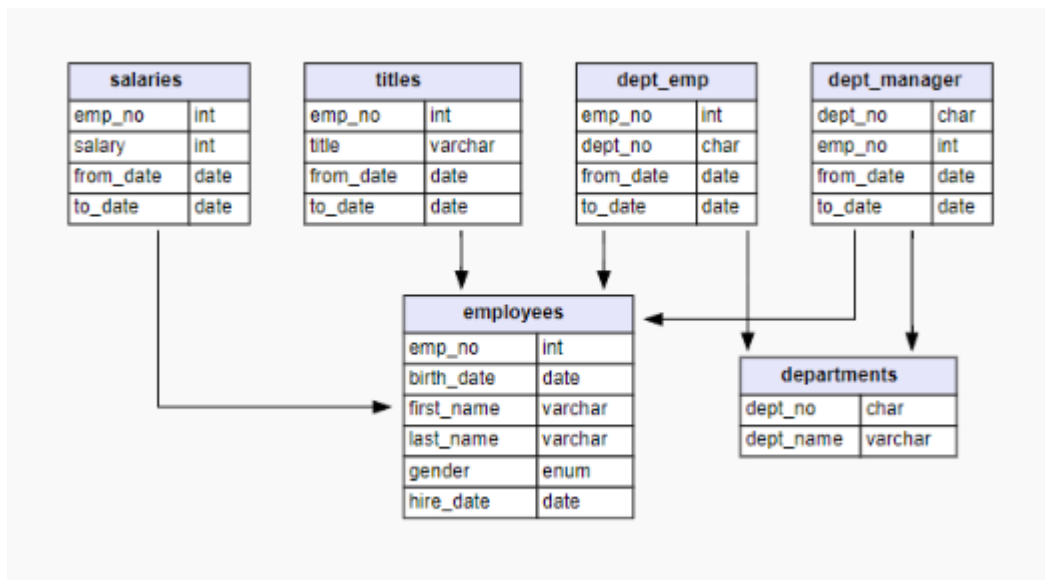


Figure 1: Database Schema

Data visualization is the activity of placing information into a visual framework, such as a map or graph, to make data simpler for the human brain to understand and draw conclusions from. The main goal of data visualization is to make it easier to see patterns, trends, and outliers in large data sets. Statistics graphics, information visualization, and information graphics are all phrases that are sometimes used interchangeably.



1.1. REASONS FOR SELECTING THE SUBJECT AREA AND DATA

Employee information is very crucial and essential for a firm where it is included with both the basic information and records of the employee history.

- It may be crucial for the company's human resource department to keep track of the following information since the dataset contains salary information for all employees, along with their department, titles, and personal information like date of birth and date of hire, although not limited to the following:
 - Age distribution among employees according to the department.
 - Corporate wage costs are broken down by Department.
 - Worker gender according to the department.

1.2. VISION AND GOALS

The main objective is to create a Data Mart for the stakeholders so they're able to view every employee that the company currently has on staff. There are two options that could be used:

Approach 1:

- Building a data warehouse and using the outputs as a starting point for developing data marts for the relevant stakeholders.
- Creating a data mart to meet the requirements of the various stakeholders' businesses.

Approach 2:

The second technique involved creating a data mart with all of the company's current employees and then filtering out the rest of the applicants. To put it another way, omitting former employees.

Data Mart makes sure that the HR Department takes into account the employees who are currently employed by the company.

- Gender and age-based segregation for a certain department.
- The distribution of assets and money given to each department.
- What budgetary allocation might each department make for the upcoming round of hiring?



High-level Decision Makers and the HR department both find it simpler to make predictions about the company's future thanks to these insights.

1.3 KEY STAKEHOLDERS

1. HR Department
2. Resource Allocation Team.
3. Decision-Making team.

1.4 BUSINESS REQUIREMENTS

1. Identify the age group that has been with the company the longest.
2. Developing a better plan of action to properly accomplish the need.
3. Establish the resource management action plan.
4. Distribution of adequate funding, according to the department, for the upcoming employment round.

2. SCHEMA / DIMENSION MODEL

Data is split into facts and dimensions in the simplest dimensional model, known as a star schema. The architecture of the relational schema in this case represents a multidimensional data model. The star schema is the explicit data warehouse schema. This schema is known as a "star schema" because its entity-relationship diagram resembles a star with points that diverge from a center table.

The motive behind the design:

- A dimensional table distinct from the calendar so that all dates, such as birthDate and hireDate, are recorded in one location and no duplicates are created. This frees up memory that would otherwise be used by redundant dates.
- A dimensional table for departments, as each employee has at least one department, and this common data may be stored separately to conserve memory.
- A fact table with simply the salary attribute, because the salary is a numerical value that changes as the person progresses in his or her career.
- A dimensional table for employee details that contains all personal



information that is unique to each individual.

Data Warehouse of Star Schema is shown below:

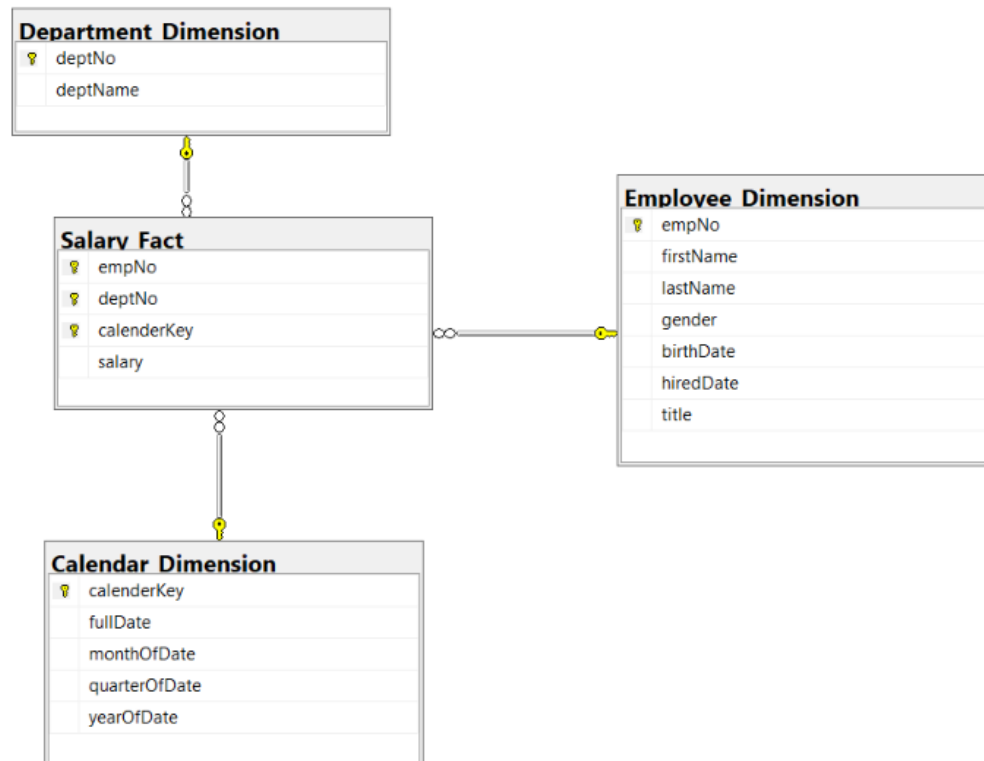


Figure 2: Star Schema for Employee Dataset



Dimensional Model of Data Warehouse is shown below:

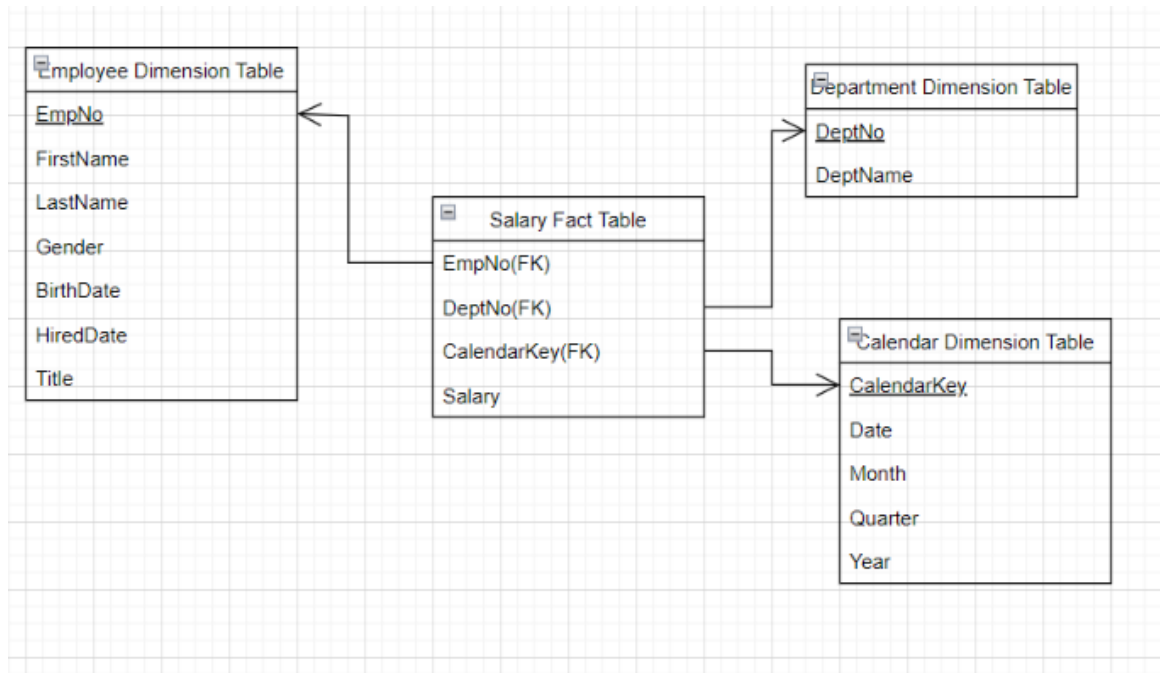


Figure 3: Dimension Model for the Dataset

In this case, the Fact table is referenced by three primary keys that were derived from dimension tables. Salary is the relevant property.

3. IMPLEMENTATION OF DATA WAREHOUSE

SQL Server Integration Service (SSIS) provides a simple and standardized way to read data from numerous sources, carry out aggregations and transformations, and then integrate the data for data warehousing and analytics. For processing large amounts of data, SSIS emerges as the ideal technique (GBs or TBs).

Prerequisites :

- Visual Studio 2019 installed
- MS SQL Server installed

Using SQL queries on MS Server, the Dimensions and fact table shown in Figure 1 are built in the database. For a list of all SQL queries for creating tables, please see the appendix.



Data is moved from the source dataset file to the target tables when tables are generated using SSIS's Data flow activities.

3.1 ETL TO POPULATE DATA WAREHOUSE

The data flow performs a special function for the control flow. The data flow has a separate tab, right next to the control flow, because it needs its canvas. The data flow is a structure that enables you to read data from various sources into the memory of the machine that is executing the SSIS program. Once the data is in memory, you can make a variety of changes to it. Because they are stored in memory, they are very rapid. After the modifications, the data is written to one or more destinations (a flat file, an Excel file, a database, etc.).

1. Employee_Dimension
 2. Department_Dimension
 3. Calendar_Dimension
 4. FactSalary
1. Employee_Dimension
We sorted the single column and combined the data for the employee_dimension table.

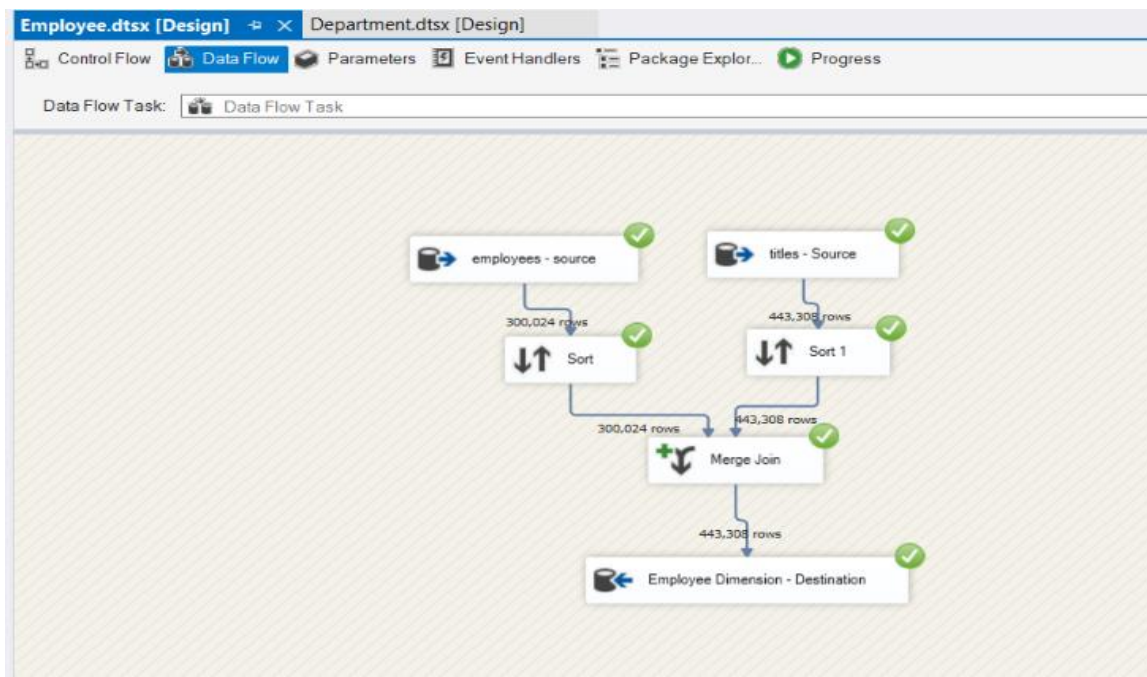


Figure 4: Employee Dimension



2. Department_Dimension

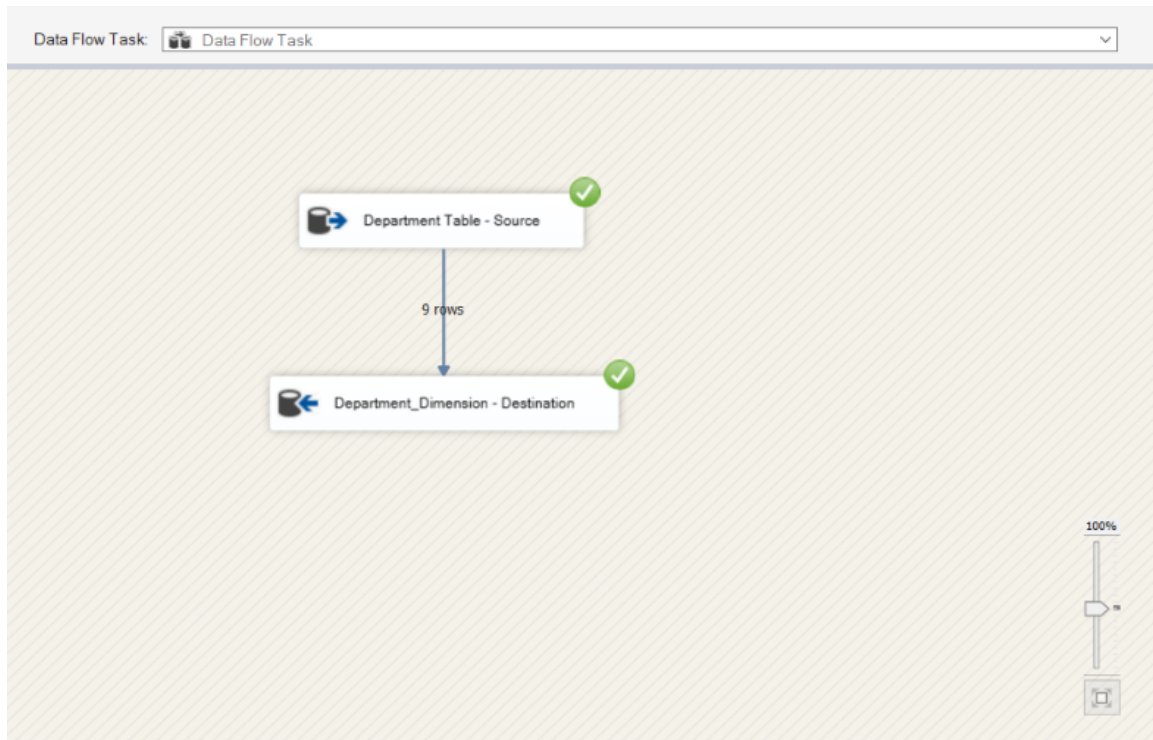


Figure 5: Department Dimension

3. Calendar_Dimension

It was required to make sure that the calendar dim table only included unique dates. So, to filter out unique dates, the hiredDate and birthDate columns were merged using a left join after being sorted using a sorting tool.



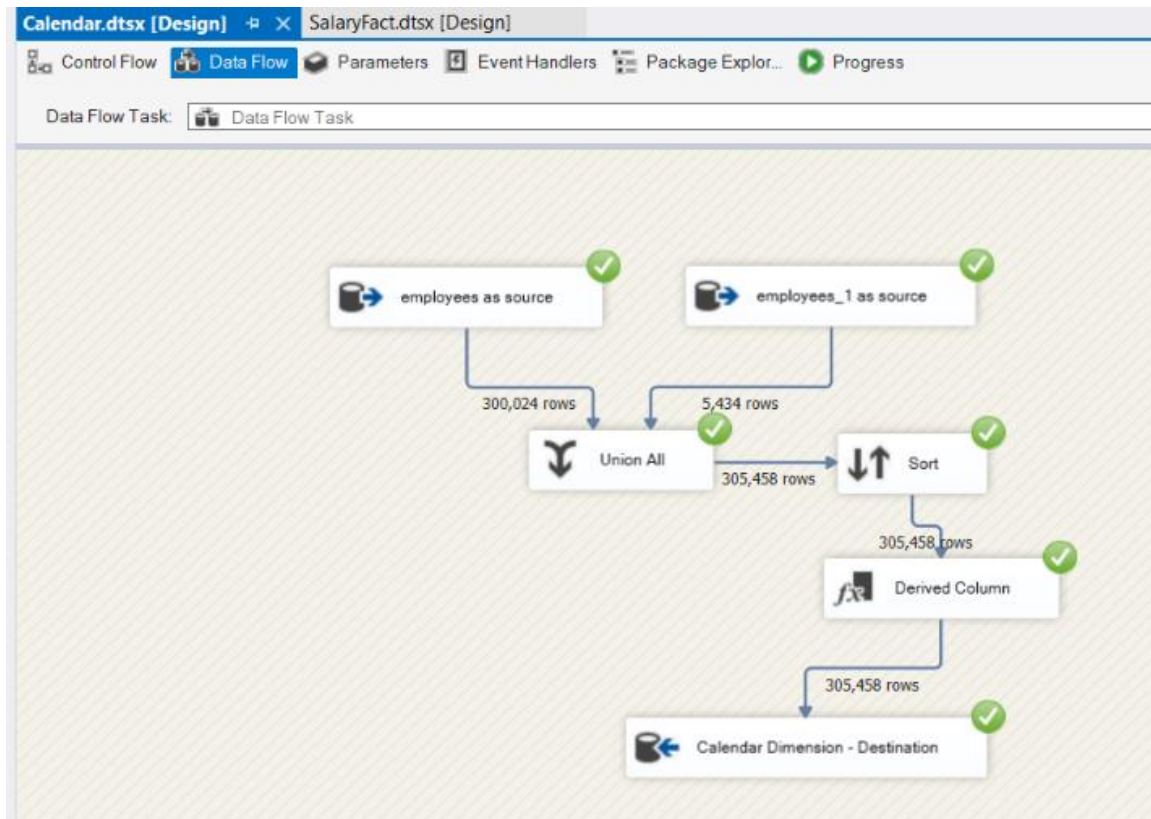


Figure 6: Calendar Dimension

Likewise, we had to make sure that the FactSalary table only included data on people who were working for the company.



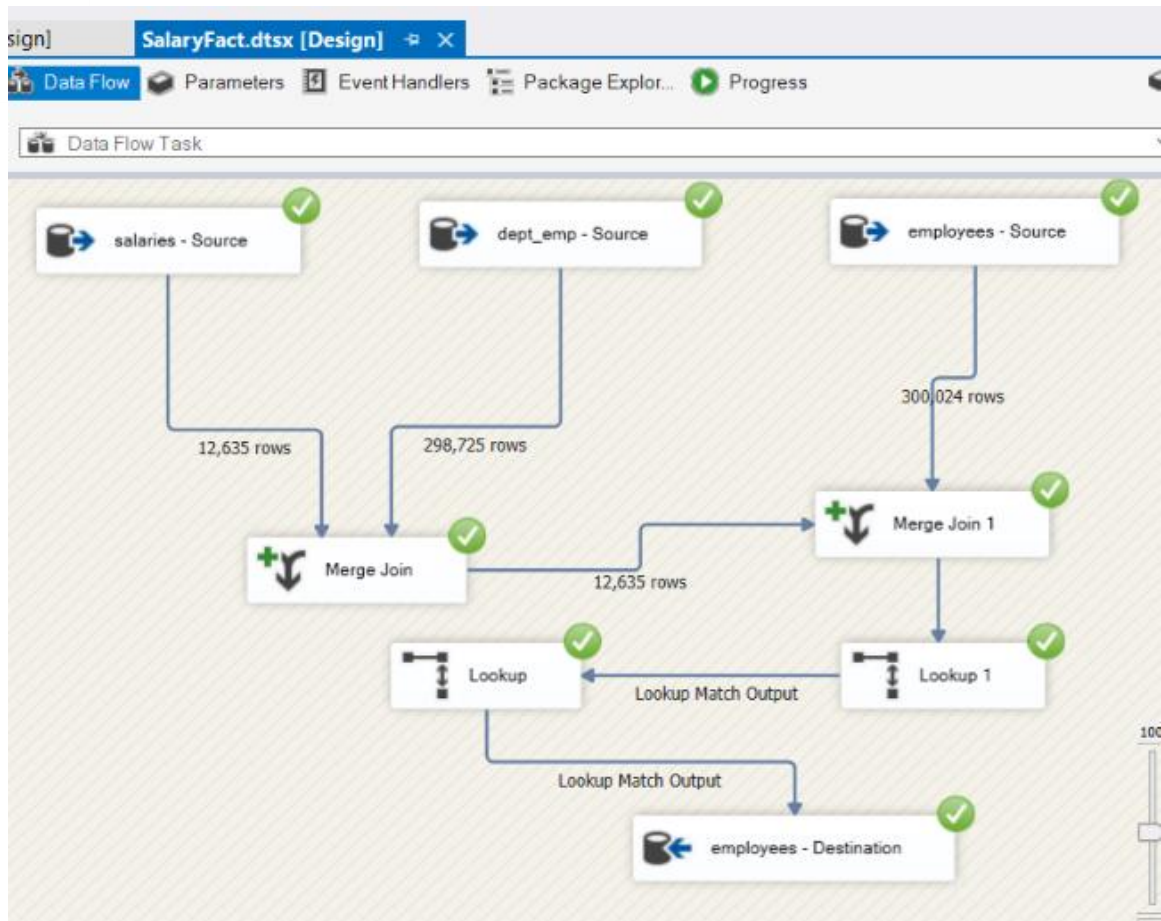


Figure 7: Fact Salary Dimension

4. VISUALIZATIONS AND REPORTS

Tableau for data visualization offers innovative ways to significantly improve the ability to uncover hidden skills. To encourage consumer interaction with the data, visual analytics are used. Using Tableau, users can interact visually with data to quickly obtain insights and make crucial decisions.

Here, we have established a connection between Tableau and MS SQL Server and have retrieved the data that I was able to upload into the dimension and fact tables.



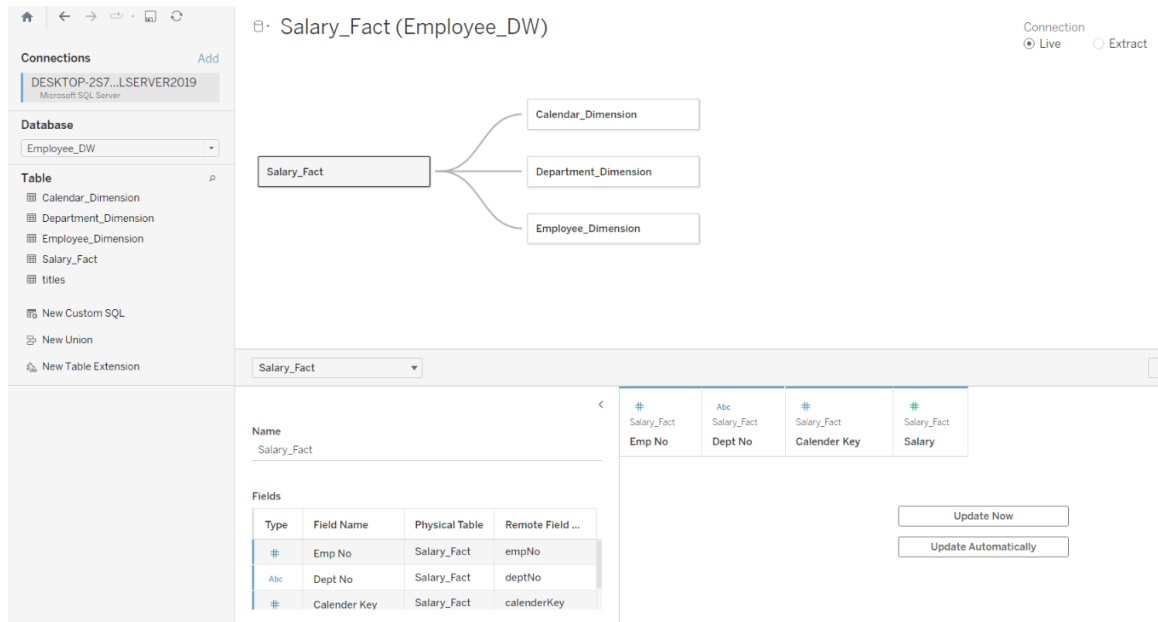


Figure 8: Database connection in Tableau

4.1 VISUALIZATIONS

Visualization 1:

1. Yearly starting salary: How much has it increased between 1985 and 2002?
 - During the years, the salary has increased by 20k (\$52k to 72k).



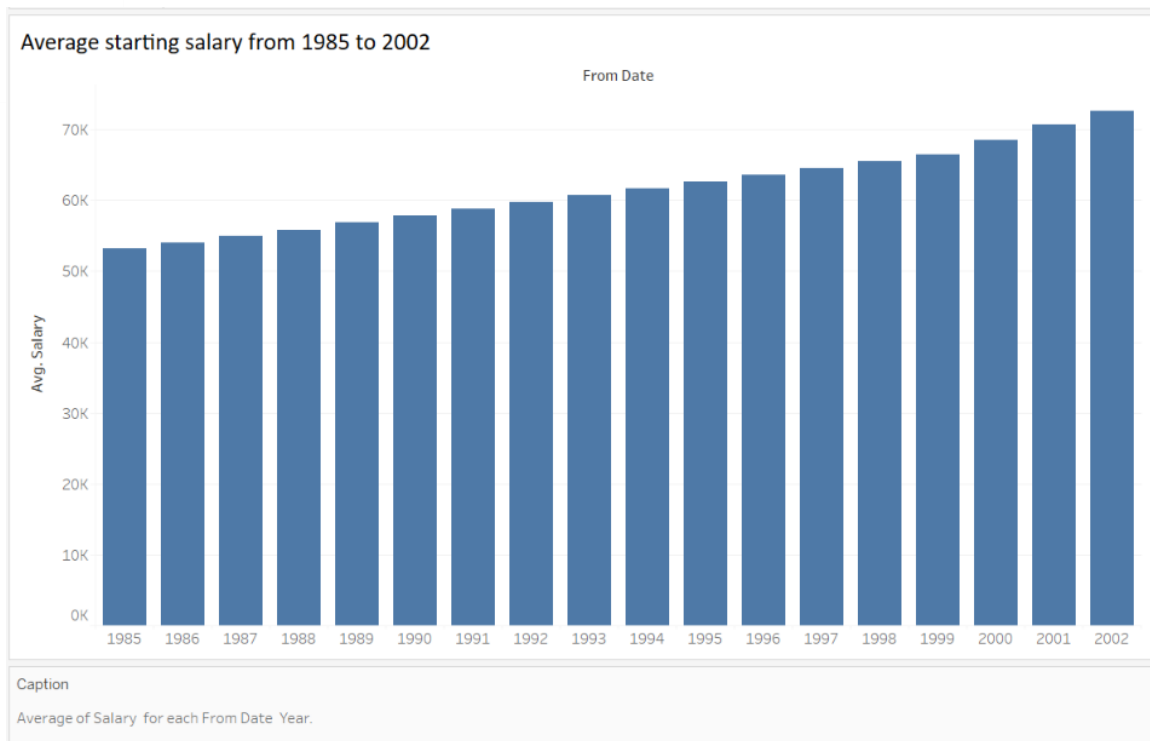


Figure 9: Average Salary in Years

Visualization 2:

2. Number of employees per job role

- Represents the number of workers in each employment role

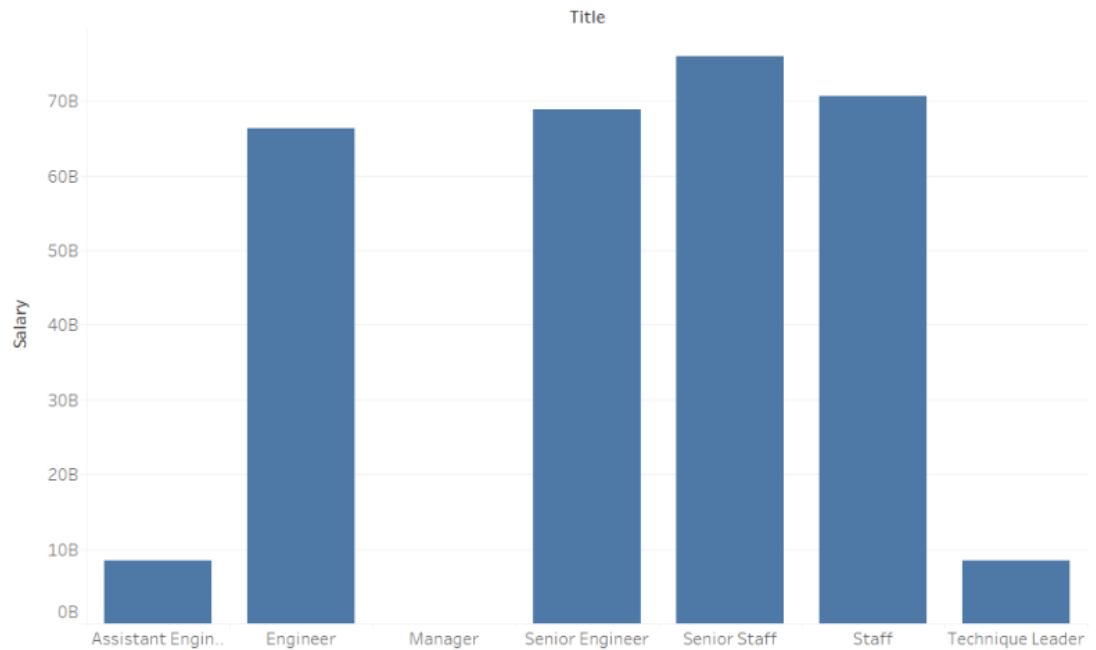


Figure 10: Number of employees per job role



Visualization 3:**3. Average salary as per Role.**

Average salary as per Role.

**Caption**

Sum of Salary for each Title.

Figure 11: Average Salary per Role**Visualization 4:****4. SD (Standard deviation) of salary as per role**

- The salary difference between various roles.



Standard deviation of salary for each Title

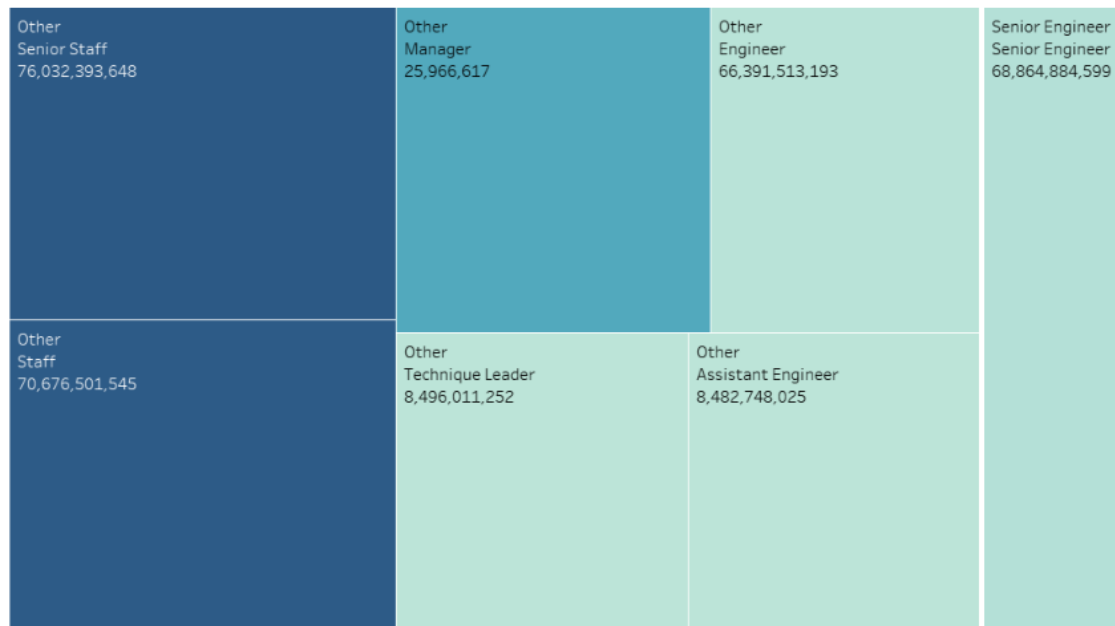


Figure 12: Standard deviation of salary per role

Dashboard:

On a dashboard, all of your data is presented visually. Although it has several uses, its primary function is to swiftly convey information, such as KPIs.

The data for a dashboard frequently originates from a linked database and is shown on a separate page. It may often be customized so that you can choose the data you want to display and whether or not you want to include charts or graphs to illustrate the data.



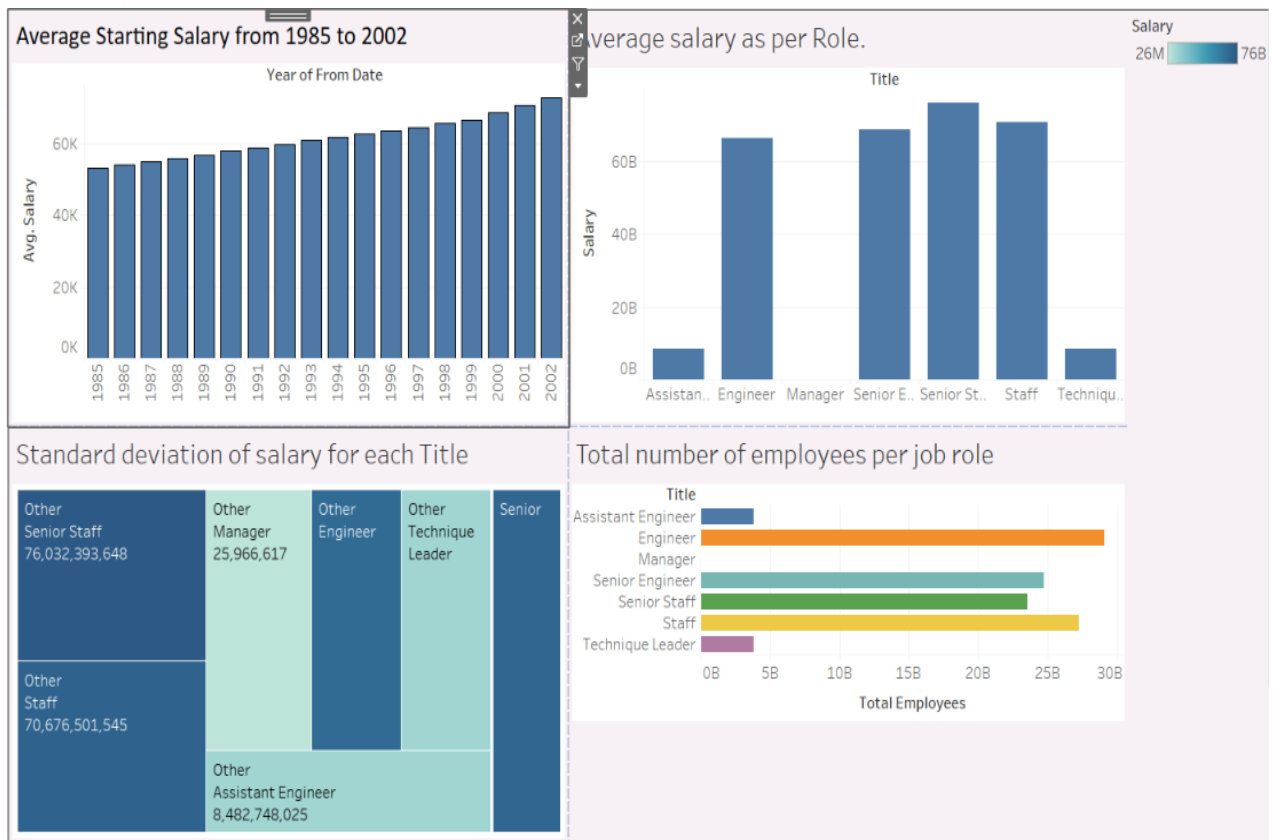


Figure 13: Dashboard

4.2 REPORTS

1. The salary distribution of Department Gender
 - Describes the gender-based distribution of the workforce and the total remuneration paid to each of those groups.



DEPARTMENT-GENDER WISE SALARY

Department Name	Gender	Count	Salary
"Customer Service"	F	7568	€ 483,508,345
	M	11486	€ 737,908,576
"Development"	F	27679	€ 1,764,433,934
	M	41963	€ 2,681,692,120
"Finance"	F	5608	€ 359,115,458
	M	8487	€ 542,108,231
"Human Resources"	F	5816	€ 370,411,153
	M	8632	€ 552,413,068
"Marketing"	F	6439	€ 412,031,481
	M	9815	€ 628,486,582
"Production"	F	23808	€ 1,518,388,188
	M	35729	€ 2,274,227,736
"Quality Management"	F	6630	€ 422,844,554
	M	9710	€ 620,764,214
"Research"	F	7020	€ 447,585,591
	M	10123	€ 646,928,472
"Sales"	F	16933	€ 1,081,283,359
	M	25493	€ 1,629,295,044

Figure 14 : Department Gender Salary

2. The report of Department Manager

- Information on each department's manager.
- The number of workers who report to the management.



DEPARTMENT WISE EMPLOYEE COUNT

Dept No	Department Name	Full Name	Count
d009	"Customer Service"	Zhongwei Czap	19054
d005	"Development"	Guiseppe Gonthier	69642
d002	"Finance"	Apostol Shihab	14095
d003	"Human Resources"	Duri Pramanik	14448
d001	"Marketing"	Guiseppe Gonthier	16254
d004	"Production"	Elvia Welham	59537
d006	"Quality Management"	Kendra Fortenbacher	16340
d008	"Research"	Satyanarayana Marchesini	17143
d007	"Sales"	Elvia Welham	42426

Figure 15: Department Employee Count

3. The report of Yearly Salary
 - Pay distribution by department, as well as information on annual changes since joining the company.



DEPARTMENT WISE SALARY BREAKDOWN

Year	Dept No	Dept Name	Total Salary Paid
1952	d001	"Marketing"	€ 1,040,518,063
	d002	"Finance"	€ 901,223,689
	d003	"Human Resources"	€ 922,824,221
	d004	"Production"	€ 3,792,615,924
	d005	"Development"	€ 4,446,126,054
	d006	"Quality Management"	€ 1,043,608,768
	d007	"Sales"	€ 2,710,578,403
	d008	"Research"	€ 1,094,514,063
	d009	"Customer Service"	€ 1,221,416,921
1953	d001	"Marketing"	€ 6,140,906
	d002	"Finance"	€ 5,408,161
	d003	"Human Resources"	€ 5,428,771
	d004	"Production"	€ 24,201,994
	d005	"Development"	€ 24,672,150
	d006	"Quality Management"	€ 5,938,373
	d007	"Sales"	€ 15,813,606
	d008	"Research"	€ 6,179,094
	d009	"Customer Service"	€ 7,925,542
1954	d001	"Marketing"	€ 8,132,795
	d002	"Finance"	€ 7,018,245
	d003	"Human Resources"	€ 5,744,230
	d004	"Production"	€ 27,833,689
	d005	"Development"	€ 31,677,846
	d006	"Quality Management"	€ 7,400,713
	d007	"Sales"	€ 18,972,336
	d008	"Research"	€ 7,974,792
	d009	"Customer Service"	€ 7,995,726
1955	d001	"Marketing"	€ 1,650,614
	d002	"Finance"	€ 1,610,697
	d003	"Human Resources"	€ 1,843,687
	d004	"Production"	€ 4,063,001
	d005	"Development"	€ 6,827,906
	d006	"Quality Management"	€ 946,045

Figure 16: Department Wise Salary



4. Department Title distribution

- Allocation of all roles and titles within each department, together with the salaries paid for each title.

DEPARTMENT TITLE WISE SALARY

Department Name	Title	Count	Salary
"Customer Service"	"Assistant Engineer"	628	€ 41,521,617
	"Engineer"	4955	€ 317,348,382
	"Senior Engineer"	4239	€ 271,122,931
	"Senior Staff"	4078	€ 261,882,286
	"Staff"	4760	€ 304,688,798
	"Technique Leader"	665	€ 42,551,706
"Development"	"Assistant Engineer"	2359	€ 151,383,054
	"Engineer"	18401	€ 1,174,509,306
	"Senior Engineer"	15638	€ 998,188,737
	"Senior Staff"	14905	€ 950,998,138
	"Staff"	16937	€ 1,080,795,292
	"Technique Leader"	2398	€ 153,429,429
"Finance"	"Assistant Engineer"	526	€ 33,829,748
	"Engineer"	3765	€ 239,141,245
	"Senior Engineer"	3119	€ 199,491,439
	"Senior Staff"	2981	€ 191,187,009
	"Staff"	3458	€ 222,023,514
	"Technique Leader"	463	€ 29,587,837
"Human Resources"	"Assistant Engineer"	530	€ 33,866,120
	"Engineer"	3778	€ 239,326,982
	"Senior Engineer"	3259	€ 209,164,498
	"Senior Staff"	3031	€ 193,039,702
	"Staff"	3564	€ 229,077,304
	"Technique Leader"	493	€ 31,366,303
"Marketing"	"Assistant Engineer"	585	€ 37,410,145
	"Engineer"	4237	€ 272,477,343
	"Senior Engineer"	3761	€ 242,119,373
	"Senior Staff"	3431	€ 216,894,562
	"Staff"	3940	€ 251,611,604
	"Technique Leader"	558	€ 35,929,351
"Production"	"Assistant Engineer"	2008	€ 128,881,413

Figure 17: Department Title Wise Salary



5. CONCLUSION

This project provides following insights:

1. The usage of SSIS and SSRS to build data warehouses and reports.
2. Tableau reports provide you more freedom to play with the data.

I learned about Microsoft SQL Server Integration Services [SSIS], associated data replication methods, and SSIS data integration through this assignment. Because many crucial components can be simply integrated, automated integration with your data warehouses/multiple data sources and the analytics database can greatly assist your decision.

As we created SSRS reports using the SSRS report builder in practice. Because of how easy it is to use, the report builder shines out and is more useful in self-service BI systems.

Users of Tableau may make complex graphs and charts that resemble pivot table graphs quickly and easily. Tableau is a robust and easy-to-use program. Tableau makes managing large volumes of data easier and provides faster dataset calculations.

6. REFERENCES:

- [1] Converting Backup files into csv files: <https://blog.devart.com/how-to-export-sql-server-data-from-table-to-a-csv-file.html> [Accessed: 22nd Feb 2022]
- [2] Data Source: <https://www.sqlskills.com/sql-server-resources/sql-server-demos/> [Accessed : 2nd Mar 2022]
- [3] Data Warehouse: <https://www.oracle.com/ie/database/what-is-a-data-warehouse/> [3rd Mar 2022].
- [4] SQL Server Management Studio: Available in system.
- [5] SSIS: Merge Join: <https://www.tutorialgateway.org/merge-join-transformation-in-ssis/>



7. APPENDIX

SQL Queries

```
create database Employee_DW
go
```

```
create table Calendar_Dimension
( calenderKey int not null identity,
fullDate datetime, --from employees table :
hire_date and birth_date
monthOfDate char(10),
quarterOfDate char(2),
yearOfDate int,
PRIMARY KEY (calenderKey)
);
go
```

```
create table Employee_Dimension
( empNo int not null identity,          --from employees table
firstName nvarchar(50),                --from employees table
lastName nvarchar(50),                 --from employees table
gender nvarchar(50),                   --from employees table
birthDate date,                        --from employees table
hiredDate date,                        --from employees table
title nvarchar(50),                    --from titles table WITH RECENT
to_date
PRIMARY KEY (empNo)
);
Go
```

```
create table Department_Dimension
(deptNo nvarchar(255) not null ,
deptName nvarchar(50),
PRIMARY KEY (deptNo)
);
go
```




```

create table Salary_Fact
(
    empNo int null,                --from DimEmp
    deptNo nvarchar(255),          --from DimDept
    calenderKey int null,          --from DimCalender
    salary int,                    --from salaries table WITH
    RECENT to_date
    PRIMARY KEY (empNo,deptNo,calenderKey),
    foreign key (empNo) references Employee_Dimension(empNo),
    foreign key (deptNo) references Department_Dimension(deptNo),
    foreign key (calenderKey) references Calendar_Dimension(calenderKey),
);
go

```

```

create table titles
(empNo int not null identity,
title nvarchar(50),
from_Date date,
to_Date date,
PRIMARY KEY (empNo)
);
go

```

Report SP's

Report :Department Gender Salary

```

USE [Employee_DW]
GO

```

```

/***** Object: StoredProcedure [dbo].[GetEmployeeDetails]   Script Date: 3/10/2023
8:53:50 PM *****/
SET ANSI_NULLS ON
GO

```

```

SET QUOTED_IDENTIFIER ON

```



GO

CREATE PROCEDURE [dbo].[GetEmployeeDetails]

AS

BEGIN

SET NOCOUNT ON

SELECT

Department_Dimension.deptName,Employee_Dimension.gender,Employee_Dimension.
empNo,Salary_Fact.salary,Department_Dimension.deptNo

FROM Salary_Fact INNER JOIN

Department_Dimension ON Salary_Fact.deptNo = Department_Dimension.deptNo INNER
JOIN

Employee_Dimension ON Salary_Fact.empNo = Employee_Dimension.empNo

END

GO

Report :Department Employee Count

USE [Employee_DW]

GO

/***** Object: StoredProcedure [dbo].[GetDeptEmployeeCount] Script Date:
3/12/2023 2:36:14 PM *****/

SET ANSI_NULLS ON

GO

SET QUOTED_IDENTIFIER ON

GO

CREATE PROCEDURE [dbo].[GetDeptEmployeeCount]

AS

BEGIN

SET NOCOUNT ON

SELECT Department_Dimension.deptName, Employee_Dimension.empNo,



```
Employee_Dimension.firstName, Employee_Dimension.lastName,  
Department_Dimension.deptNo  
FROM Department_Dimension INNER JOIN  
Salary_Fact ON Department_Dimension.deptNo = Salary_Fact.deptNo INNER JOIN  
Employee_Dimension ON Salary_Fact.empNo = Employee_Dimension.empNo  
  
END  
GO
```

Report: Department Employee Count

```
USE [Employee_DW]  
GO
```

```
/***** Object: StoredProcedure [dbo].[GetDeptEmployeeCount] Script Date:  
3/12/2023 2:36:14 PM *****/
```

```
SET ANSI_NULLS ON  
GO
```

```
SET QUOTED_IDENTIFIER ON  
GO
```

```
CREATE PROCEDURE [dbo].[GetDeptEmployeeCount]  
AS  
BEGIN  
SET NOCOUNT ON
```

```
SELECT Salary_Fact.deptNo, Calendar_Dimension.yearOfDate, Salary_Fact.salary,  
Department_Dimension.deptName  
FROM Calendar_Dimension INNER JOIN  
Salary_Fact ON Calendar_Dimension.calenderKey = Salary_Fact.calenderKey  
INNER JOIN  
Department_Dimension ON Salary_Fact.deptNo = Department_Dimension.deptNo
```



END
GO

Report: Department Title Wise Salary

USE [Employee_DW]
GO

/***** Object: StoredProcedure [dbo].[GetDeptEmployeeCount] Script Date:
3/12/2023 2:36:14 PM *****/

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE PROCEDURE [dbo].[GetDeptEmployeeCount]
AS
BEGIN
SET NOCOUNT ON

SELECT Salary_Fact.deptNo, Salary_Fact.salary, Department_Dimension.deptName,
Employee_Dimension.empNo, Employee_Dimension.gender, Employee_Dimension.title
FROM Salary_Fact INNER JOIN
Department_Dimension ON Salary_Fact.deptNo = Department_Dimension.deptNo INNER
JOIN
Employee_Dimension ON Salary_Fact.empNo = Employee_Dimension.empNo

END
GO

