PROJECT

IS 594 Data Warehousing

Name: Vismay Bhavinkumar Patel

Date: 05/19/21

Company background and Business Objective:

There is a fictious company **George Healthcare**, which is working with the state government closely to monitor the healthcare system to leverage the diagnosis of medical condition, disease or any other sickness. For example, cancer, liver tumour, viral infections, etc. This project initiated by the collaboration of the government and private organization, can foster to achieve advancement in healthcare and research, treat and help people to lead a healthy lifestyle.

The objective of this program is to help the doctors and health centres to identify pattern in recognizing unknown virus or diseases or medical conditions, by keeping track of symptoms associated with existing diseases/infections/virus/medical conditions. Please note that here the term disease, infections, virus or medical condition are used in the same reference to the problem statement interchangeably.

Keeping the track of symptoms allows the doctors to diagnose the disease early and faster. Also, it allows the doctors to easily distinguish between the symptoms caused by different medical conditions. Generally, the treatment of common flue or viral infections are easy, because they happen during season change and are identified with less effort. However, other sever infections or viral conditions like pneumonia, tuberculosis, appendix or liver cancer are not easy to identify. It is because the symptoms for these conditions build slowly or start appearing in the body with time, this can lead to misunderstanding or unclear diagnosis, resulting in prescribing different drugs or medicine which belong to the same family of drugs.

Let me give an example of fungal infection, fungal infection generally effects people during summers due to humid conditions and sweating, or they can also spread to someone, if they come in contact with surface infected with fungus capable to infect humans. As a result of the infection, the basic symptoms are irritation or itchy skin, rashes, red ness of the skin white yeast like bacteria accumulating in the skin etc.

The medicine prescribed for fungal infection varies based on severity levels, the family of medicine that are generally prescribed for fungal infection are “Azole” clotrimazole, miconazole, Fluconazole, itraconazole etc. These are types of antifungal creams or tablets that the doctor prescribes, coming back to the problem domain, if the patient is not able to communicate the proper symptoms or if the doctor fails to diagnose the severity of the infection, the initial condition can get sever, since the infection can return back when medication with insufficient concentration of medicine is prescribed.

How this will help us to encourage people to adopt healthy lifestyle? When the patients come up with different medical condition or infections, after the diagnosis and start of the treatment, doctors can help the patients by providing guidelines, that they need to follow or things to avoid for the similar condition or similar type of medical condition to return. As we discussed in over example, the precautions that the patient can take in case of infections is to keep the surrounding clean, avoid accumulation of water at corners in the house, take bath twice a day during summers and properly drying the body, because fungus resides on surfaces with humid and warm surfaces.

Hence, the objective defined by George Healthcare and the state government's collaboration can help the healthcare facilities and organizations to combat sever and more complex medical conditions or diseases, harmful to human health and survival.

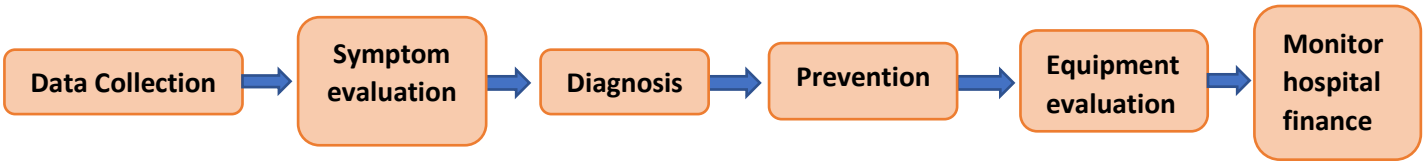
Problems / gaps addressed by DW/BI at overall organization:

Following are the problems addressed at organizational level:

- Diagnosis of a major medical condition at early stage, resulting in excessive use of resource and time.
- Need support of technology to bolster solution to analyse problems.
- Provide better healthcare treatment and consulting/therapy based on the severity of the medical condition. i.e., if the doctors have facts and cause of problem and a possible solution to the problem. We can help the patients to be mentally strong not give up.
- Need to initiate a step towards advancement of medical and healthcare facilities provided by healthcare institutes and organizations.

Process Maps:

Business Process:



In the Process map / business process consists of 6 processes, as shown above. Followed by the business process is the table of Business Process – Activity – KPI – Facts – Dimension. This explains the activities that take place in each process and the Key Performance Indicators of the processes.

Based on the KPI's the facts and dimensions of the dimensional model are assumed, so that a clear understanding can help design a lucrative database using dimensional modeling where the data will be stored in the later process.

Connecting Business Process, KPI and Dimensions:

Business Process	Activity	KPI	Facts	Dimensions
Data Collection	- Survey. - Interviews. - Identify data. structured data. unstructured data. -collect data in file/text.	-n% of population suffer from specific disease/condition. -Z number of male / females diagnosed with X infection. -Yearly turnover of hospital. -Salary of doctors based on expertise. -Alkaline levels in patient.	Total Population Total population by gender. Yearly Turnover Salary Alkaline level	By city/state By Gender By Hospital/asset funds By Doctor name Patient name/gender.
Symptom evaluation	-Allergy / reactions on skin. -Smoking history. -Evaluate Allergy. -MRI scans. -Take blood test. -Any case of cold, indigestion, infection, weakness.	Presence of tumour in organ. % Of fat in arteries. chance of infectious disease, exposure. Organ functioning issues, if any. Possibility of stroke or hypertension. Brain disorder.	Tumour Size. Fat percent. Steroid / alkaline level in liver. BP reading. Damage % in Brain.	By Age. By arteries (Left / right) By Yes/No By High/ Moderate /Low By Chance of Survival.
Diagnosis	-Conformation of medical condition. -make report -prescribe medication/treatment	Rate of spread of infection/disease. Patient follow up rate.	Percent of spread in Body. Number of visits for treatment.	By Body Organs. By Week
Prevention	-Taking healthy diet. -Physical activities. -Avoid drugs and alcohol. -Proper sleep and focus on Mental health.	% Of time spent on physical activity. % Of time spent on mental health. % Of reduction on consumption of Alcohol.	Hours Minutes	By days By weeks

Equipment evaluation	-keep track of equipment shortage. -maintenance / cost. -Usage of equipment in diff. departments.	-Working equipments. -Total Requirement. -Cost of equipment.	Equipment Quantity. Price Per unit.	By Product name. By Product name.
Monitoring hospital finance	-asset management -investing. -tracking profit -maintaining cash records.	-Asset Investment. -Capital Growth. -Annual Turnover	Asset amount. Liquid cash amount. Annual Growth amount.	By Category of asset.

Figure 0.1

Common Dimensions:

Business Problem	Patient	Symptoms	Disease Diagnosis	Treatment	Equipment
Data Collection	X	X		X	
Medication	X	X		X	
Advance Research		X	X		X
Better treatment		X	X	X	X
Faster diagnosis of disease.	X	X			X
Identification of distinct and unique parasites.		X	X		X
Preparedness form outbreaks of disease.		X	X		

Figure 0.2

Dimensional Model diagram:

The dimensional model diagram consists of database design, and relation between the tables, represented by foreign key and primary key. This design can further be elaborated into fact and dimension tables, fact tables generally collect the performance measures generated by an organization's activities. The values in the fact table are numeric, additive, semi-additive. The fact table always expresses a many to many relationships, and it is made up of the foreign keys of the intersecting dimensional model. A dimensional table is one which is uniquely identified by a single key, that is a unique numeric value mostly integer. The below dimensional model is a replica of the database tables built on MySQL Workbench Database Model as an Entity relationship Diagram.

Moving further let us explore the Fact tables and dimension tables separately:

First Fact table Equipment:

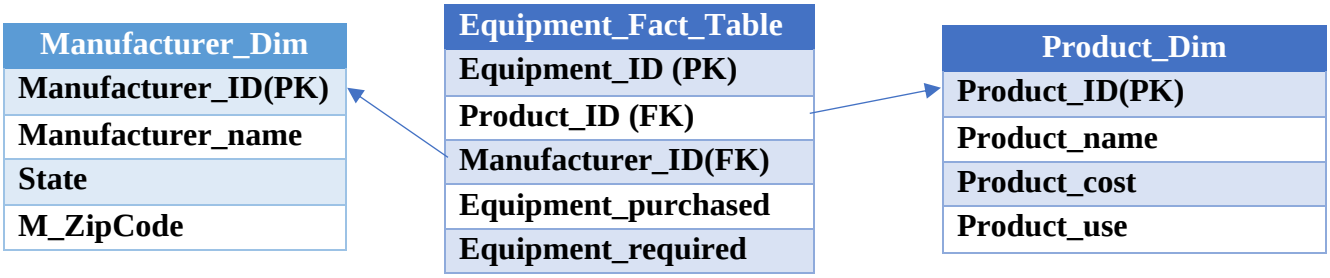


Figure 0.3

The above three tables contain two-dimensional table and one fact table, namely product and manufacturing, these tables contain equipment information used in the hospitals. Manufacturer dimension contains the manufacturer details like Manufacturer_ID, Manufacturer name, state and zip code for location. Product dimension contains product name, product cost, total units, product use. Both the dimension tables are connected to the fact table equipment as foreign keys. The equipment fact table consists of equipments in the hospital, i.e., the product, manufacturer, requirement and number of units purchased.

Second Fact Table Report:

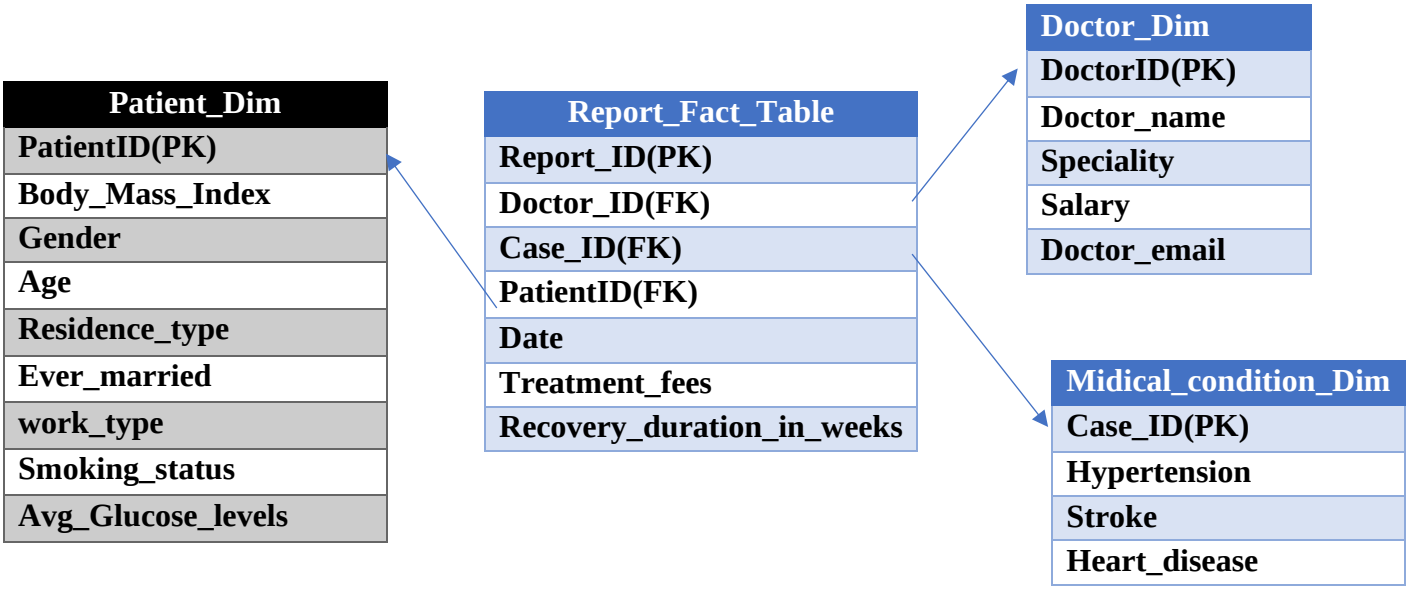


Figure 0.4

The above tables contain information regarding medical condition of the patient, doctor and report. Patient, Doctor and medical condition tables are dimension tables, connected to report fact table as foreign key. Hypertension, stroke, heart_disease are the dimensions which contain numeric values, we know that dimension tables only contain non-numeric values, but when the values do not change frequently a numeric dimension can be considered in the dimensional model. The same is the case with medical condition levels, any person diagnosed with the above medical condition, takes time to recover and the change in the values is not frequent. For example, a specific level of hypertension or stroke needs to be observed, and monitored, because we only get a stroke or hypertension if the levels cross certain threshold.

Third Fact table Hospital:

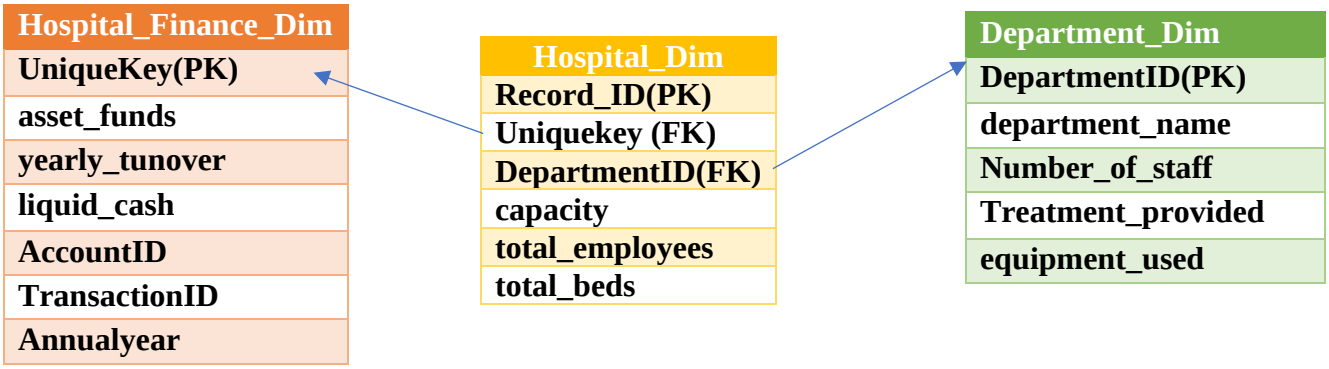


Figure 0.5

The table consists of hospital information regarding the departments and hospital finance. Department and hospital finance are dimensional tables, and hospital is a fact table connected to them with a foreign key. The fact table contains information of hospital capacity, total_employees and number of beds. Just like the previous table the hospital finance has numeric values, since we are monitoring the finances annually the values are not changing frequently by month so we can consider it in the dimensional table.

Fourth Fact Table:

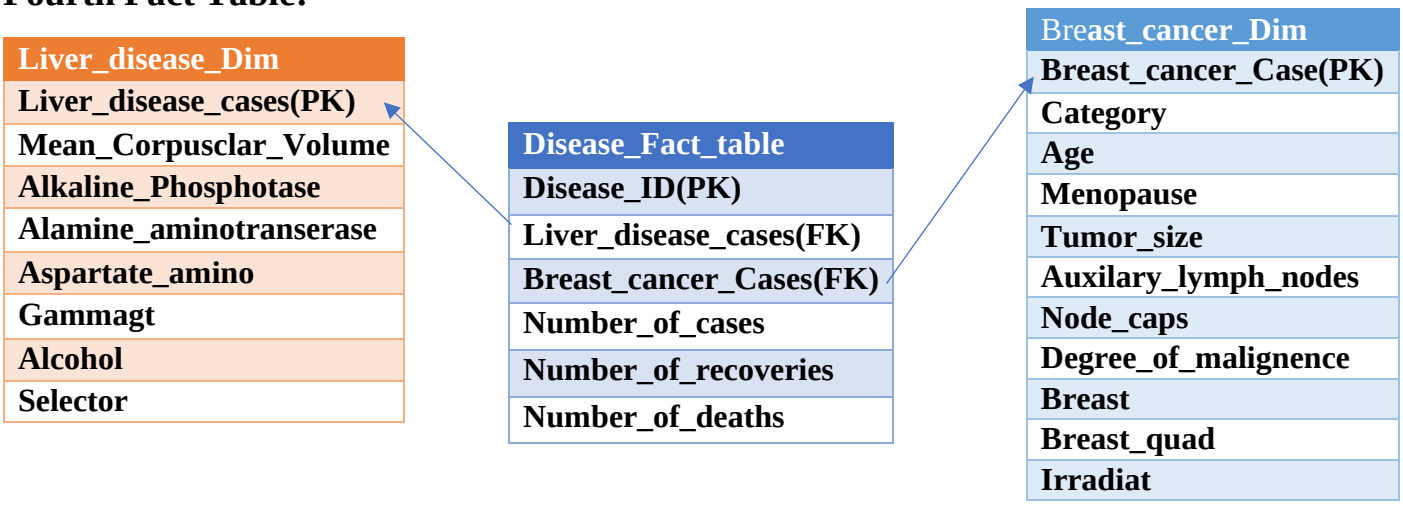


Figure 0.6

The above fact table consists of two dimensions Breast_cancer and Live_disease, which are foreign keys to disease table. Liver_disease table contains numeric values and is also a dimensional table, since the levels of liver chemicals keep changing and it is not used for calculation purpose, we can consider it in a dimensional table. Mean Corpusclar volume is a test used to check the size of the blood cells, if the size of the cells are smaller than normal then there is deficiency in the blood cells which can lead to iron deficiency, and other problems leaving impact on liver. However, the dimensions in Breast cancer are complete categorical values, because the age range and size of tumour are expressed in range of size, for example 10-20, etc. Lastly disease fact table consists of number of cases, recoveries and deaths encountered.

Now, we will look at the diagram of dimensional model in the database model that we made in MySQL workbench.

Fact tables:

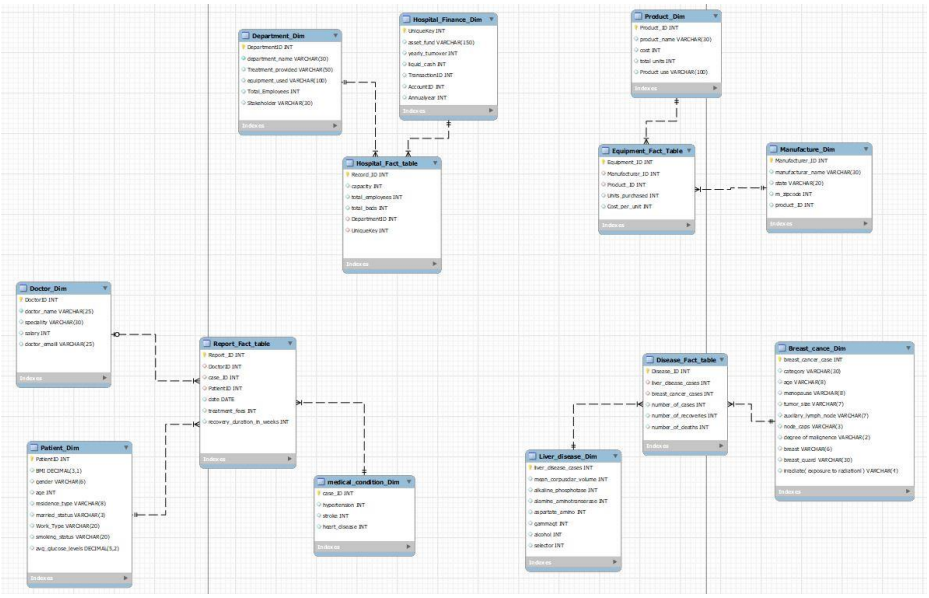


Figure 0.7

Star schema with main aggregate fact table in middle:

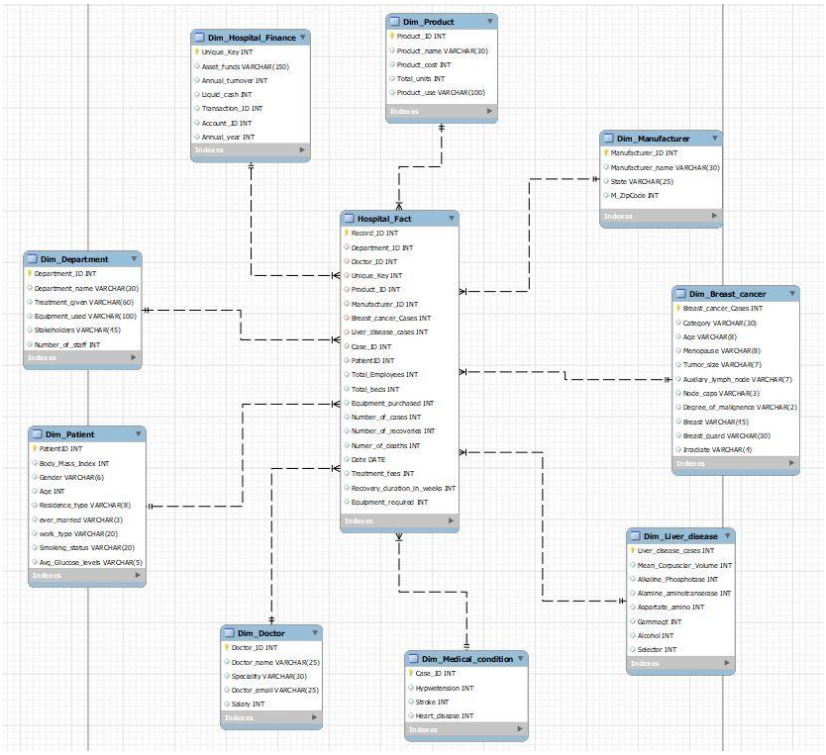
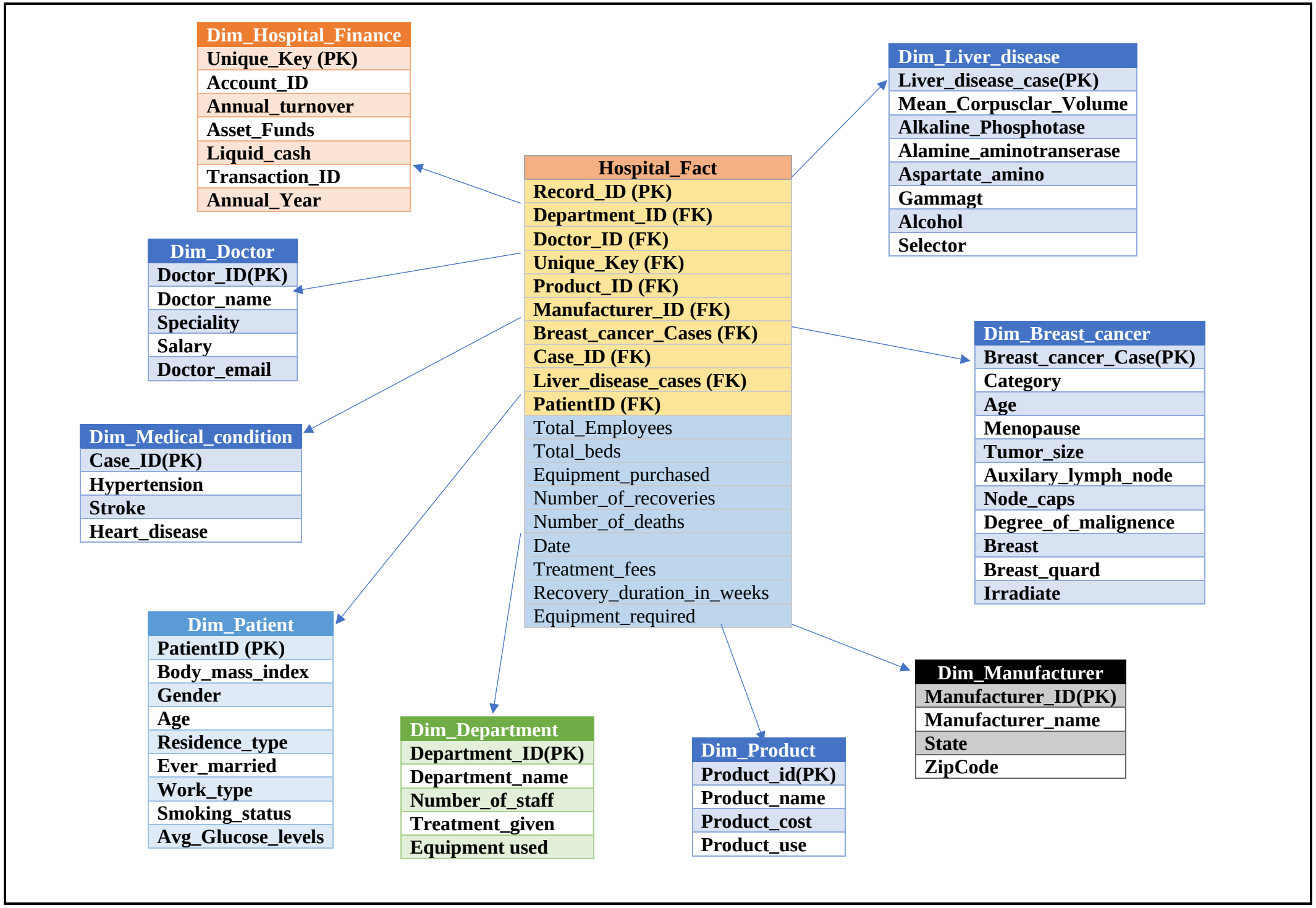
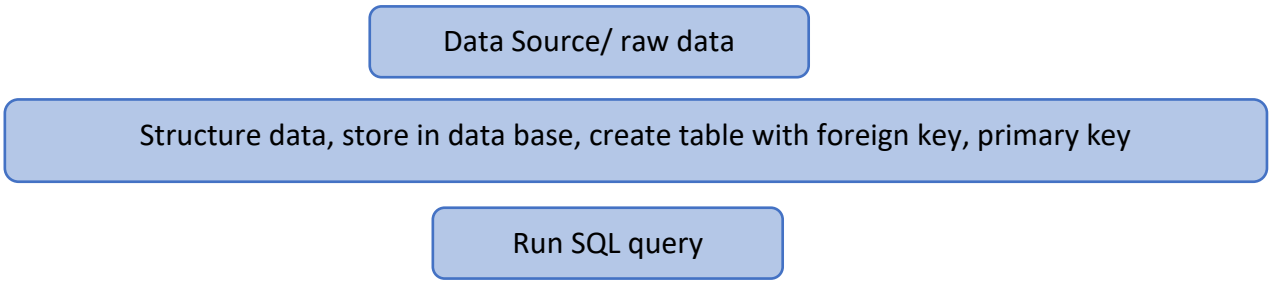


Figure 0.8

The star schema is generated by dimension tables and one aggregate fact table, the fact table contains foreign keys to dimensions and contains aggregated facts from other fact tables. The fact table consists facts measures of hospital, and foreign keys from department, hospital finance, doctor, patient, manufacturer, product, liver_disease, breast_cancer and medical condition. The fact table has one to many relation with the dimension tables, in the database design.


Dim_Hospital_Finance
Unique_Key (PK)
Account_ID
Annual_turnover
Asset_Funds
Liquid_cash
Transaction_ID
Annual_Year
Dim_Liver_disease
Liver_disease_case(PK)
Mean_Corpusclar_Volume
Alkaline_Phosphotase
Alamine_aminotranserase
Aspartate_amino
Gammagt
Alcohol
Selector
Hospital_Fact
Record_ID (PK)
Department_ID (FK)
Doctor_ID (FK)
Unique_Key (FK)
Product_ID (FK)
Manufacturer_ID (FK)
Breast_cancer_Cases (FK)
Case_ID (FK)
Liver_disease_cases (FK)
PatientID (FK)
Total_Employees
Total_beds
Equipment_purchased
Number_of_recoveries
Number_of_deaths
Date
Treatment_fees
Recovery_duration_in_weeks
Equipment_required
Dim_Doctor
Doctor_ID(PK)
Doctor_name
Speciality
Salary
Doctor_email
Dim_Breast_cancer
Breast_cancer_Case(PK)
Category
Age
Menopause
Tumor_size
Auxilary_lymph_node
Node_caps
Degree_of_malignence
Breast
Breast_quard
Irradiate
Dim_Medical_condition
Case_ID(PK)
Hypertension
Stroke
Heart_disease
Dim_Patient
PatientID (PK)
Body_mass_index
Gender
Age
Residence_type
Ever_married
Work_type
Smoking_status
Avg_Glucose_levels
Dim_Department
Department_ID(PK)
Department_name
Number_of_staff
Treatment_given
Equipment used
Dim_Product
Product_id(PK)
Product_name
Product_cost
Product_use
Dim_Manufacturer
Manufacturer_ID(PK)
Manufacturer_name
State
ZipCode

Dimensional Model Query processing:




```
67 •  SELECT Asset_Funds, Annual_turnover, Annual_year FROM data_warehousing.dim_hospital_finance ORDER BY Annual_year;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Asset_Funds	Annual_turnover	Annual_year
MANUFACTURING, TRANSPORT	200000000	2010
MANUFACTURING, TRANSPORT	246000000	2012
MANUFACTURING, TRANSPORT, STOCK	298203000	2014
MANUFACTURING, TRANSPORT, STOCK, REAL-ESTATE	350000000	2016
MANUFACTURING, TRANSPORT, STOCK, REAL-ESTATE	490234000	2018
MANUFACTURING, TRANSPORT, REAL-ESTATE	100238000	2020
MANUFACTURING, TRANSPORT, REAL-ESTATE, TECHLOLOGY	520000000	2022

Figure 0.9

Selecting Queries/KPI from Hospital Finance system table by using SQL/MySQL

The above SQL query shows monitoring annual trunover by Asset funds, over the span of 10 years. Here, we are monitoring the growth and variation observed in the annual turnover of hospitals every 2 years. Just to make the observations realistic the COVID pandamic, is kept into accunt, and prediction of growth by FY 2022 is being assumed/prodicted.

```
48 •  SELECT Department_name, Equipment_used, Number_of_staff FROM data_warehousing.dim_department;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Department_name	Equipment_used	Number_of_staff
Neurology Department	Gas anesthesia system, Inotophoresis stimulator, etc	20
Cardiology Department	Patient monitor, Cath Lab, Balloon Pump, Heart-Lung Bypass	25
Radiology Department	X-Ray, CT-Scan, MRI, Radio-Isotope scan, NMR	10
Laboratory	test kits, lab glassware, gloves, thermostats, microscopes	30
Epidemiology/Infection control	PPE kits, Sterializer, Ultraviolate light, liquid chemicals	40
Ophthalmology	Auto refractor, Retinal camera, Binocular Loupes, Trial frames	10
Psychiatry and Psychology	CT-scan, MRI, Brain imaging, EEG, Psychotherapy, etc	15
Gastroenterology Department	Ultrasound, liver biopsy, needle biopsy, MRI, CT-scan	35

Figure 1.0

Selecting KPI Query for Equipment mentioned in Department table.

Another, KPI assessment shown here is the number of equipments used per department, and the number of employees required or needed for the smooth functioning of the departments. When the hospitals are busy skilled workers, help to mitigate waiting timings, and sufficient number of equipments help to operate faster.

```
312 •  SELECT * FROM Manufacturar LEFT JOIN Equipment_detail ON Manufacturar.Manufacturar_ID = Equipment_detail.Manufacturar_ID;
313
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Manufacturar_ID	Company_name	State	M_ZipCode	Equipment_ID	Manufacturar_ID	Product_ID	Cost	Total_units	Requirement
12300234	CareFusion	Chicago	60698	10	12300234	10118	2000	12	8
12300234	CareFusion	Chicago	60698	13	12300234	10115	5000	25	30
12300234	CareFusion	Chicago	60698	23	12300234	10124	450000	5	3
12300234	CareFusion	Chicago	60698	26	12300234	10123	250000	25	20
12301122	Alaris	California	90011	11	12301122	10116	500	40	30
12301122	Alaris	California	90011	21	12301122	10125	72000	10	8
12301244	CRITICARE	Indiana	46012	19	12301244	10256	12000	20000	10000
12301345	RESPIRONICS	Chicago	60618	20	12301345	10212	700	40000	30000
12302243	Burdick	Texas	75001	12	12302243	10117	3500	20	15
12303397	Baxter	Chicago	60608	14	12303397	10113	100	400	250
12303397	Baxter	Chicago	60608	22	12303397	10114	1250	6000	6500
12304567	Bovie	Florida	32004	15	12304567	10233	10000	15	10
12304567	Bovie	Florida	32004	24	12304567	10122	20000	1000	1010
12305498	ACCU-SCOPE	California	91331	16	12305498	10111	150	50000	30000
12307888	AMSCO	Mississippi	38620	17	12307888	10112	1500	1000	900
12307888	AMSCO	Mississippi	38620	25	12307888	10214	15000	2500	2480
12309245	Bair Hugger	Buston	2112	18	12309245	10222	400	90000	80000

Figure 1.1

Performing Left JOIN on the Equipment and Manufacturing and product tables

The table above is a joint query between equipment, manufacturer, and product tables. Which shows the manufacturer name, product ID, cost of each product, total requirements of the equipments and the number of units purchased by the hospital.

```
29 •  SELECT Doctor_name, speciality, Doctor_email FROM data_warehousing.dim_doctor WHERE Doctor_email LIKE 'JA%';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Doctor_name	speciality	Doctor_email
Jack Barns	Cardiac Surgen	JABA@HOSP.org
James Potter	Critical care	JAPO@HOSP.org

Figure 1.2

The above table consists of doctor's name starting from A using LIKE operator.

The above query shows the doctors information, i.e., doctor's name, speciality and contact information. This query also highlights that, the administration staff can access specific information about the doctor using SQL query LIKE operation. Example shown here is of accessing the doctor's information by contact information, which is unique to all the doctors working in the hospital

Data Quality:

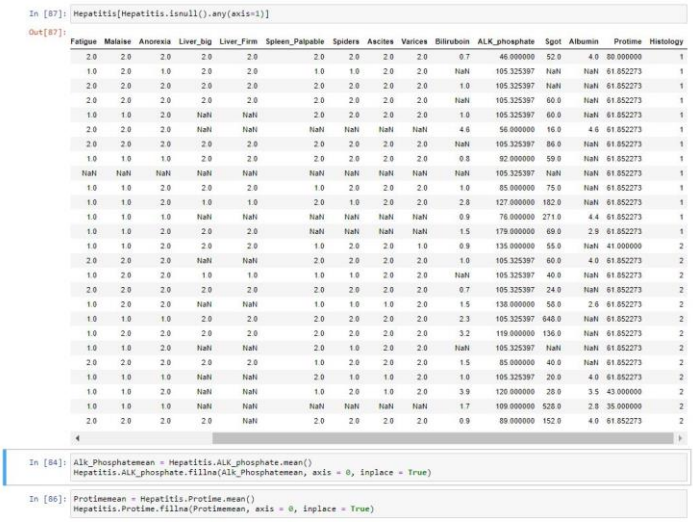
Data Quality refers to the overall utility of a dataset as a function of its ability to being processed and analysed for other users, usually by a database, data warehouse, or data analyst system.

We will list out the issues and fixes we made with the data in the over data warehouse system. Below is the table that lists the following.

ERROR CHECKING:

Serial Number	Missing values	Total rows	Total columns	Total rows after data cleaning / correction.
Breast cancer data	8	285	10	277
Liver disease	0	344	7	344
Health data	1000	5111	12	4000
Hepatitis	37	180	20	143

In out of 15 tables, we have missing values and error correction in the above four tables. Where the total rows are mentioned including missing values, Total columns, total rows and number of rows after data cleaning. Screenshots of error checking of tables are being posted of few tables, the error checking was done in python and SQL.



```
In [87]: Hepatitis[Hepatitis.isnull().any(axis=1)]
Out[87]:
```

Fatigue	Malaise	Anorexia	Liver_big	Liver_Firm	Spleen_Palpable	Spiders	Ascites	Varices	Bilirubin	ALK_phosphate	Sgot	Albumin	Protime	Histology
2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	0.7	46.000000	52.0	4.0	80.000000	1
1.0	2.0	1.0	2.0	2.0	1.0	1.0	2.0	2.0	NaN	105.325397	NaN	NaN	61.852273	1
2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	105.325397	NaN	NaN	61.852273	1
2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	NaN	105.325397	60.0	NaN	61.852273	1
1.0	1.0	2.0	NaN	NaN	2.0	2.0	2.0	2.0	1.0	105.325397	60.0	NaN	61.852273	1
2.0	2.0	2.0	NaN	NaN	NaN	NaN	NaN	NaN	4.6	56.000000	16.0	4.6	61.852273	1
2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	NaN	105.325397	86.0	NaN	61.852273	1
1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	0.8	92.000000	59.0	NaN	61.852273	1
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	105.325397	NaN	NaN	61.852273	1
1.0	1.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	1.0	85.000000	75.0	NaN	61.852273	1
1.0	1.0	2.0	1.0	1.0	2.0	1.0	2.0	2.0	2.8	127.000000	182.0	NaN	61.852273	1
1.0	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	0.9	76.000000	271.0	4.4	61.852273	1
1.0	1.0	2.0	2.0	2.0	NaN	NaN	NaN	NaN	1.5	179.000000	69.0	2.9	61.852273	1
1.0	1.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	0.9	135.000000	55.0	NaN	41.000000	2
2.0	2.0	2.0	NaN	NaN	2.0	2.0	2.0	2.0	1.0	105.325397	60.0	4.0	61.852273	2
1.0	2.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	NaN	105.325397	40.0	NaN	61.852273	2
2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	0.7	105.325397	24.0	NaN	61.852273	2
1.0	2.0	2.0	NaN	NaN	1.0	1.0	1.0	2.0	1.5	138.000000	58.0	2.6	61.852273	2
1.0	1.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.3	105.325397	648.0	NaN	61.852273	2
1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	3.2	119.000000	136.0	NaN	61.852273	2
1.0	1.0	2.0	NaN	NaN	2.0	1.0	2.0	2.0	NaN	105.325397	NaN	NaN	61.852273	2
2.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	1.5	85.000000	40.0	NaN	61.852273	2
1.0	1.0	1.0	NaN	NaN	2.0	1.0	1.0	2.0	1.0	105.325397	20.0	4.0	61.852273	2
1.0	1.0	2.0	NaN	NaN	1.0	2.0	1.0	2.0	3.9	120.000000	28.0	3.5	43.000000	2
1.0	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	1.7	109.000000	528.0	2.8	35.000000	2
2.0	2.0	2.0	2.0	NaN	2.0	2.0	2.0	2.0	0.9	89.000000	152.0	4.0	61.852273	2

```
In [84]: Alk_Phosphatemean = Hepatitis.ALK_phosphate.mean()
         Hepatitis.ALK_phosphate.fillna(Alk_Phosphatemean, axis = 0, inplace = True)

In [86]: Protimemean = Hepatitis.Protime.mean()
         Hepatitis.Protime.fillna(Protimemean, axis = 0, inplace = True)
```

Above figure is for hepatitis table consisting of null values, which are removed by taking the mean for numeric values and deleting the number of rows for categorical values, it needs to be noticed here that the data is in complete numeric format, where categorical values are described in Binary format.

```
91 •  SELECT COUNT(Node_caps) FROM Breast_cancer WHERE Node_caps = '';
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(Node_caps)
8

The above code snippet is error checking in MySQL using SQL query language, checking null values in Breast cancer data. According to the query there are 8 missing values in Node_caps column in Breast cancer data.

STANDARDIZATON:

Moving ahead with standardization, there are tables in which the standardization with respect to primary key has been done. However, the standardization has not been performed using Character values, it is done using numeric values, following are the tables where the method is performed.

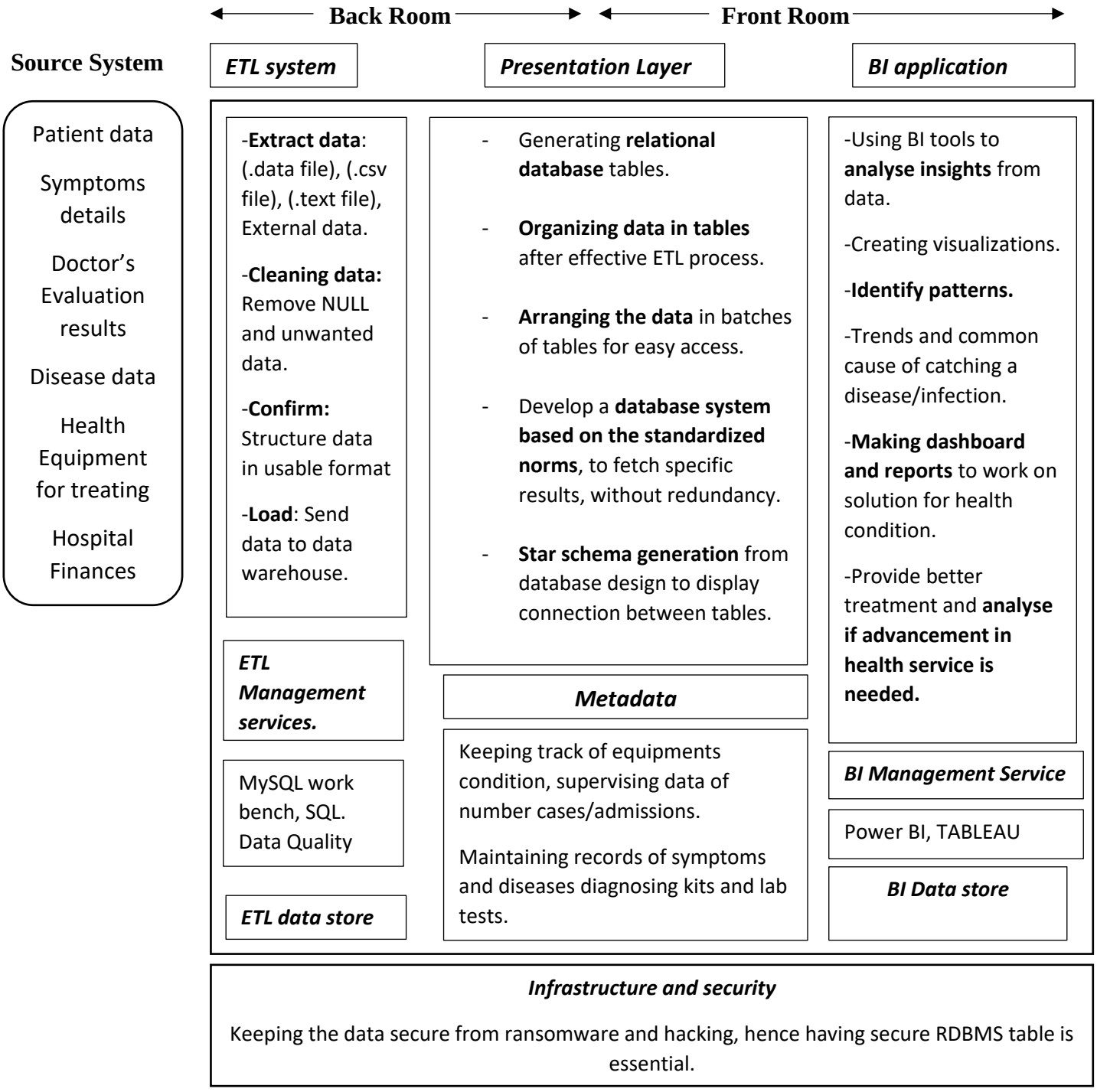
AREA OF APPLICABILITY	STANDARD	EXAMPLE
Doctor_email	The doctor's email will use first 2 letters of First name, and last name, followed by @ and hosp .org example.	Flin Rogre Email: FLRO@HOSP.org
Product name	Product name will contain first 3 capitalized letters of the company, as standard.	For example, company name: CareFusion Standard: CARtestkits
Manufacturer ID	The manufacturer ID is a 8 digit id with first 5 digits assigned to standardization norms. The first 3 digits represent country code, the next 2 digits represent state code, followed by random numbers.	**12305**498 – here 123 is the code for United States, and 05 is the state code for California State. **12302**243 – similarly 123 is the code for USA and 02 is the code for Texas state.
Equipment Detail	The equipment record in the hospital database has a similar but different standardization process. % Digit number with 0 as fix starting initial followed by 4-digit number. The series for Equipments start from 00010.	**0**0010 – unique identity, says the equipment detail from hospital record starts from 0. This entry is for equipment from CareFusion manufacturer based in Chicago **0**0019 – This entry is for CRITICARE manufacturer based in Indiana.
Hospital Finance	To maintain the integrity of the Hospital Finance account information, the access to the Transaction ID is unique to a combination of characters and numbers.	**HPLF 0012 2654 5668**

The standardization is done in 5 tables in the data warehouse system, that helps to uniquely identify the data and fetch the data in the system.

- Standardization is done to avoid repetition, faster access to unique and specific information.
- It helps to avoid misunderstandings, when we look at the data. For example, a product brought form two different companies, if the product name is same then, we need to look up the entire record for company name and their products.
- Applying Standardization eliminates addition time consumption.

ARCHITECTURE:

A. End to End Architecture:



The above diagram shows end-to-end architecture, that has two main components Back room and front room components. These components consist of source system, ETL system, presentation system and BI application, where ETL and half of presentation layer comes under back room and BI application and remaining half of presentation layer comes under front room component. The systems also contain ETL data source, metadata and BI data store, which store the data related to each component separately. At the end there is Infrastructure and security section that monitors threat to data, like cyber hacking, loss of data, ransomware, etc.

Source System:

The source system is the place from where the data comes into the data warehouse system. The data can be structured and unstructured or semi structured data, we have used both structured and unstructured data in the data warehouse, the structured data comes in .csv file, and .data file. The .data file is unstructured which only contained data points, with no column values, the data was separated by ‘,’(comma). The .csv files were structured containing missing values. The rest of the tables or data was test data, that was generated internally.

Patient details, disease details, symptoms for different diseases, this data is extracted from external sources. This includes Liver_disease, Stroke, Heart disease, Breast_cancer, Hepatitis, Patient details, which includes different symptoms or parameters by which different medical condition/disease are caused. Equipment, Doctor, Product, Manufacturer data are internally structured generated data.

ETL System;

The ETL system is based on Extracting, Transforming and Loading the data into the data warehouse or data marts, it can be enhanced further into Extracting, Cleaning, Confirm and Deliver. Here, the system first extracts the data from the data source, which is structured, unstructured and semi-structured data. Once the data is extracted and saved, in the form of text, csv, data files. Next, the data is transformed into data, which can be made ready for operational use. In my project I have done extract, clean, confirm and load manually with python programming using Jupyter notebook framework, by removing null data and cleaning it. I have also used MySQL workbench and SQL queries to transform data. There are many automated tools by which the transformation of data can be achieved, when the data is large and complex, these tools are AWS Redshift, ORACLE warehouse, IBM DB2, and more.

After removing the null values and other unwanted data, the data was ingested into SQL tables designed in MySQL using database model.

Presentation Layer:

The presentation layer is responsible to represent the relational database tables that will be generated from the data loaded after ETL process, in this stage there was a Star Schema generated using the various fact and dimensional tables.

Business Intelligence Application:

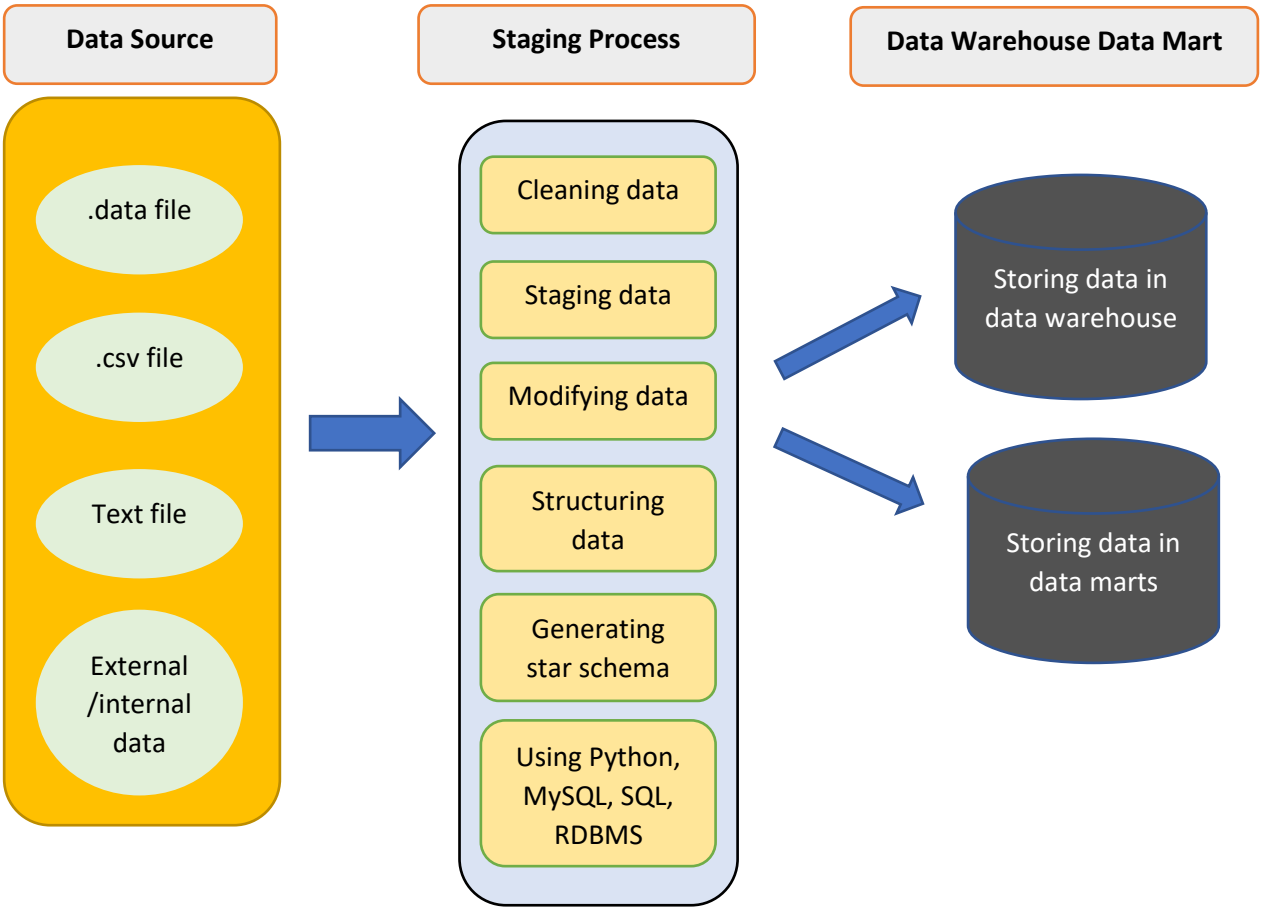
Business Intelligence deals with the visualization, analysis and graphical representation of the data from the presentation layer. In over case we will be generating dashboards and reports from of various health datasets to justify and fulfil the objectives of the problem statement. Tools used in BI application will be Power BI.

Apart from all these components there are four other components in the end-to-end architecture, ETL data store, BI data store, metadata and infrastructure and security, these components store the specific data of the designated systems. While Infrastructure and security manage and maintain the architecture of the components.

B. Source Systems Table:

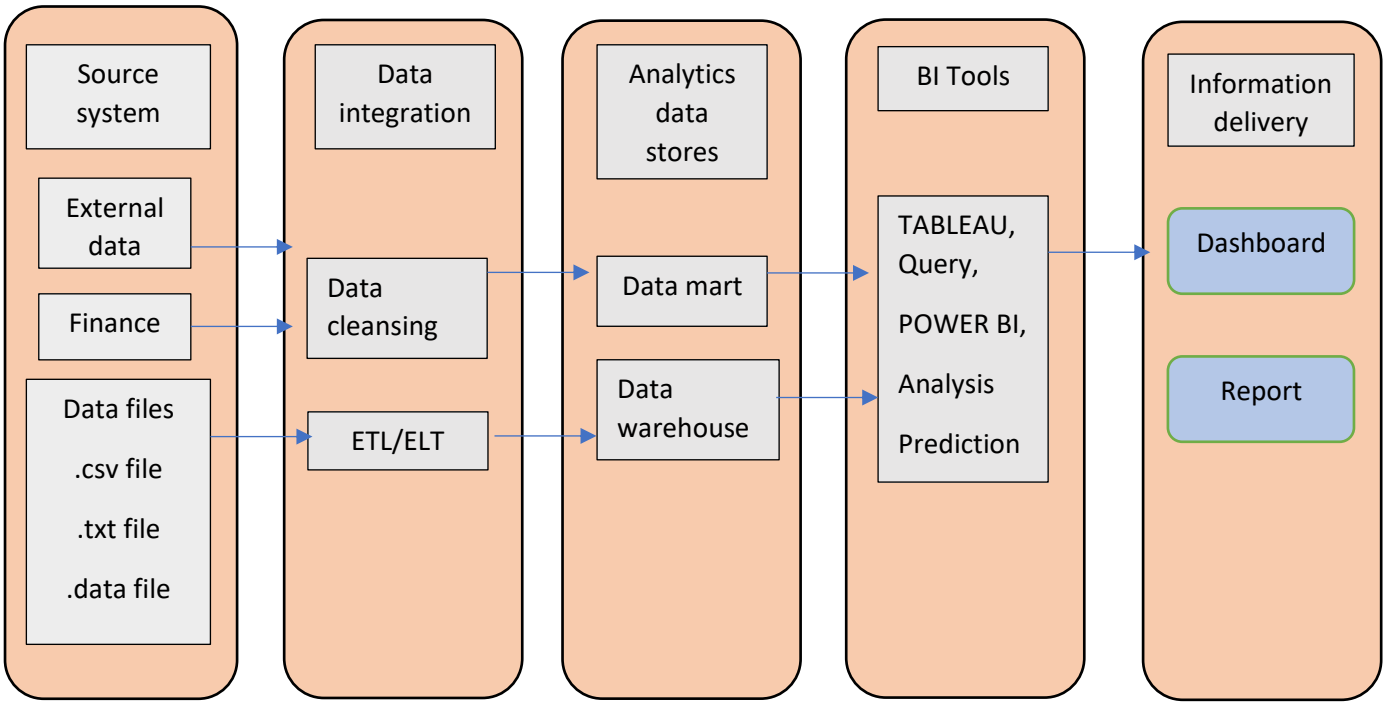
Source	Business Owner	IS Owner	Platform	Location	Description
Symptoms (External)	Josh Marsh	Martha Kane	Windows	HQ-Chicago	Categorical and numeric data of different symptoms related to diseases. Ph levels of body, from research labs and other institutes.
Doctor's evaluation report (Internal data)	Hospital Administration	None	Windows/AWS	HQ – Chicago	Doctors' evaluation reports, based on symptoms consisting of data in spreadsheet format.
Finance (External data)	Mark lui	James Petron	Unix	HQ- California	Financial details and financial turnover and money flow of healthcare facilities.
Cancer Research institute (External data)	Brian Gemmell	Laura Marce	Windows	HQ-Washington	Different Breast cancer diagnosis and symptoms.
Healthcare data repository (External)	Jane Michel	Bruce M.	Windows	HQ – Indiana	Consists of Liver disorders, data including stroke, heart disease and other medical conditions.
Equipment data (Internal)	Jack Miller	Lucy Pevancy	Unix	HQ-Chicago	Data about the medical supplies, medical equipment, laboratory tools and tests conducted in lab.
Disease data (External data)	John Martian	Clark Kent	Unix / IBM	HQ – Chicago	Consists of different Disease data that exist.

C. ETL Components:



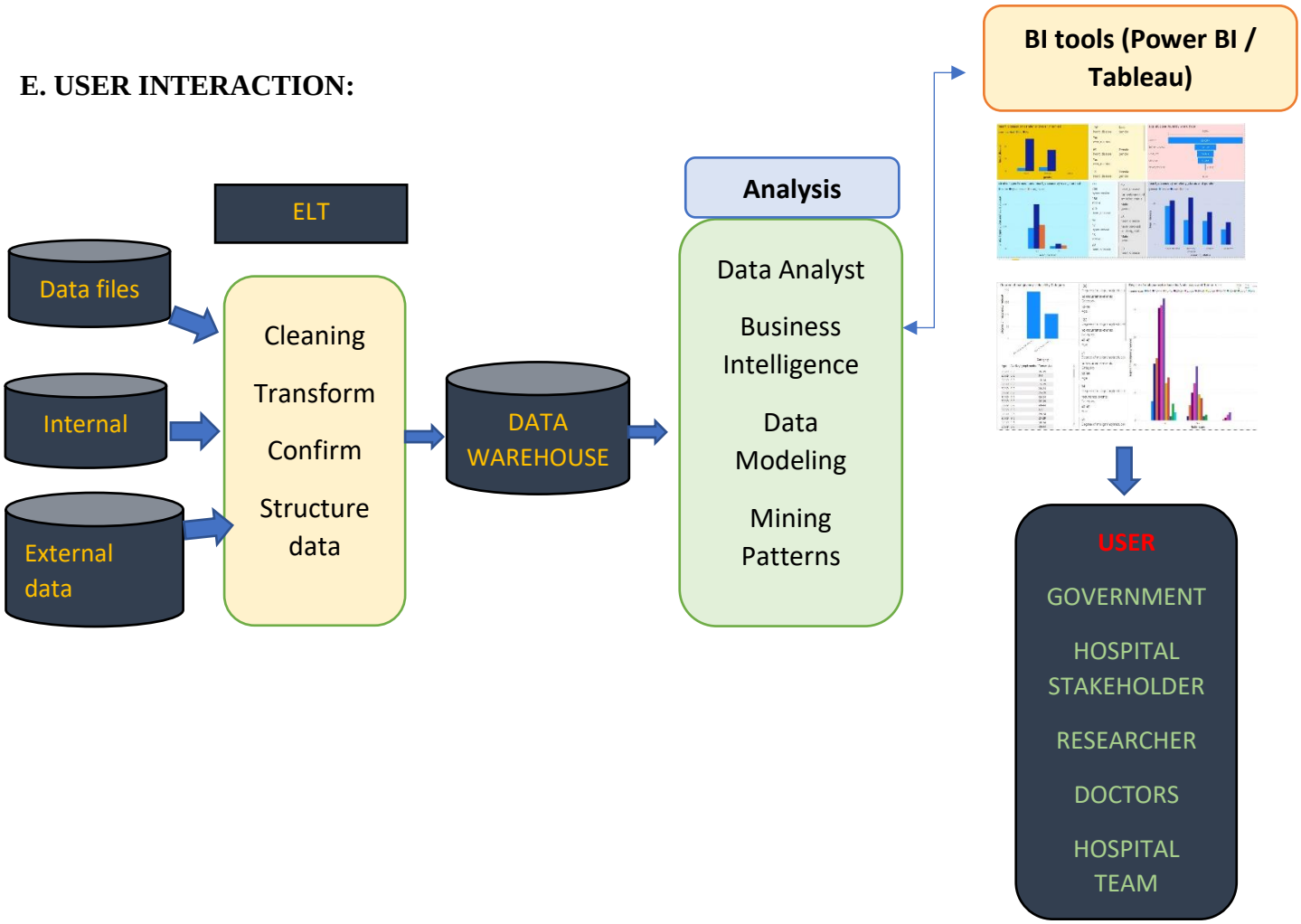
The above process describes the ETL staging where data coming from data source, in the form of .data, .csv and text file is cleaned, structured and put into proper tables. The data is organized, pre-processed and stored in to data warehouse.

D. Business Intelligence Tools:



The business intelligence tools perform analysis on the data, to represent the data coming from presentation layer/data warehouse in meaningful dashboards and reports. This helps the business user to understand the statistics and analytical information of the project with the help of visualization. Allowing the organization and business user to interact and make lucrative decisions.

E. USER INTERACTION:



The User interaction stage is where the communication between the organization and user takes place based on the product satisfaction or project objectives. The business intelligence or business analyst team creates reports and visualizations, based on analytical, mining, prediction tasks that are performed on the data, in the BI applications layer. The business user is exposed to the technical reports in the form of dashboards, charts, graphs, etc. So that they can relate to the insights generated by the business intelligence team, given that the business user might not have ace level technical background. The Business users in over project are government authorities, hospital stakeholders, researchers, doctors and hospital teams.

DASHBOARD and REPORT:

The dashboards below are visualizations from breast cancer disease.

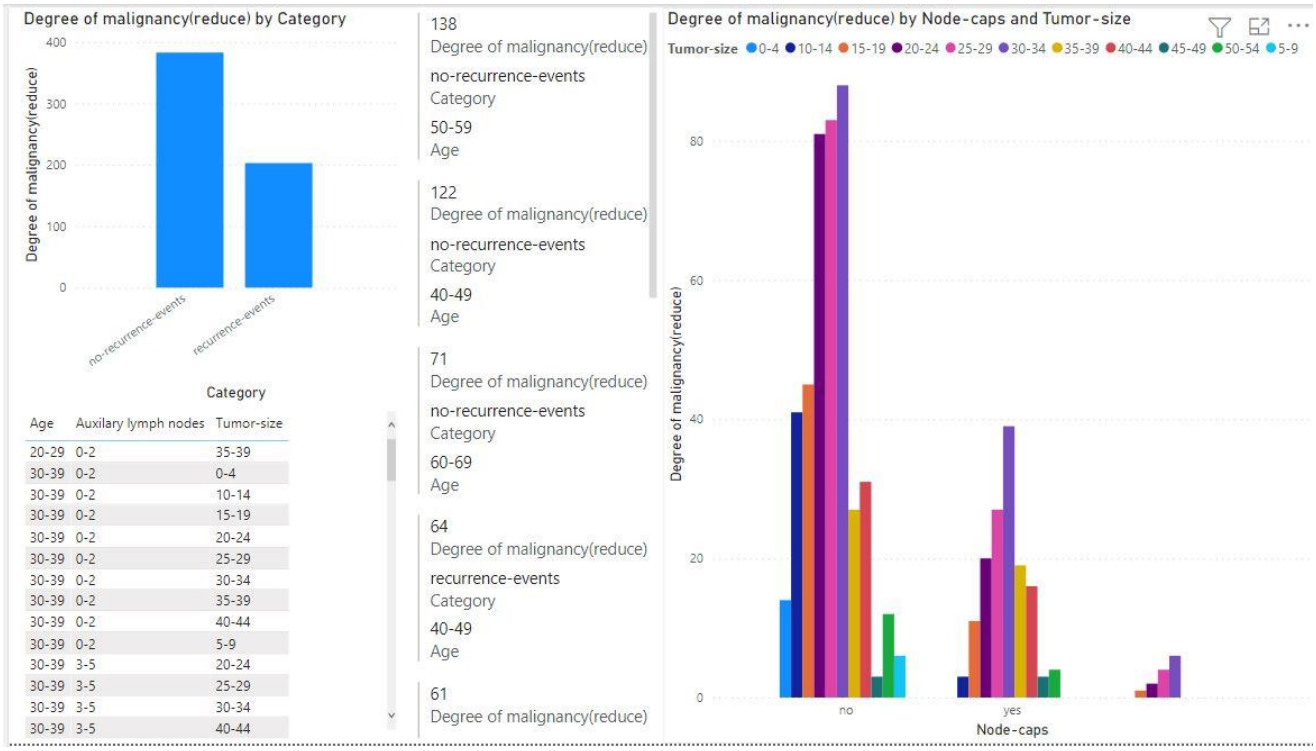


Figure 1.3

In the above dashboard shows the degree of malignancy based on category of cancer, Node-caps, Tumour-size, and reports of Auxiliary lymph nodes and Tumour size, with respect to age. The report in the middle is for the degree of malignancy based on category. First let us understand degree of malignancy, it is a term for diseases in which abnormal cells divide without control and can invade nearby tissues. Malignant cells can also spread to other parts of the body through the blood and lymph systems (germ fighting network).

Based on the above results we can make the following conclusions:

- The degree of malignancy for non-recurring events, i.e., chances of getting diagnosed with breast cancer again, is more than recurrence events, i.e., chances of getting diagnosed with breast cancer again. Also, based on the report there are more cases of recurrence events in adults below the age of 60 than adults above 60 years of age. This depicts that, adults in the range of 40-49 or 30-50 need to pay more attention on their health and the way they lead their lifestyle.
- Node-caps are caused when the cancer cells break away from the original/primary tumor, and travel through the blood or lymph system to form a new tumor in other organs or tissues in the body, such cells are covered by a cap called node caps. Based on the graph we can see that, there are more cases of breast cancer with absence of node-caps, compared to the breast-cancer cases with node-caps. However, in both the cases the tumor-size between 20-34 have higher degree of malignancy, which means they divide faster without control as compared to the size of other cancer cells.

The below dashboard is the continuation of breast cancer disease, showing degree of malignance effects on the patients based on different factors like radiation exposure, age, tumor size and Axillary lymph nodes.

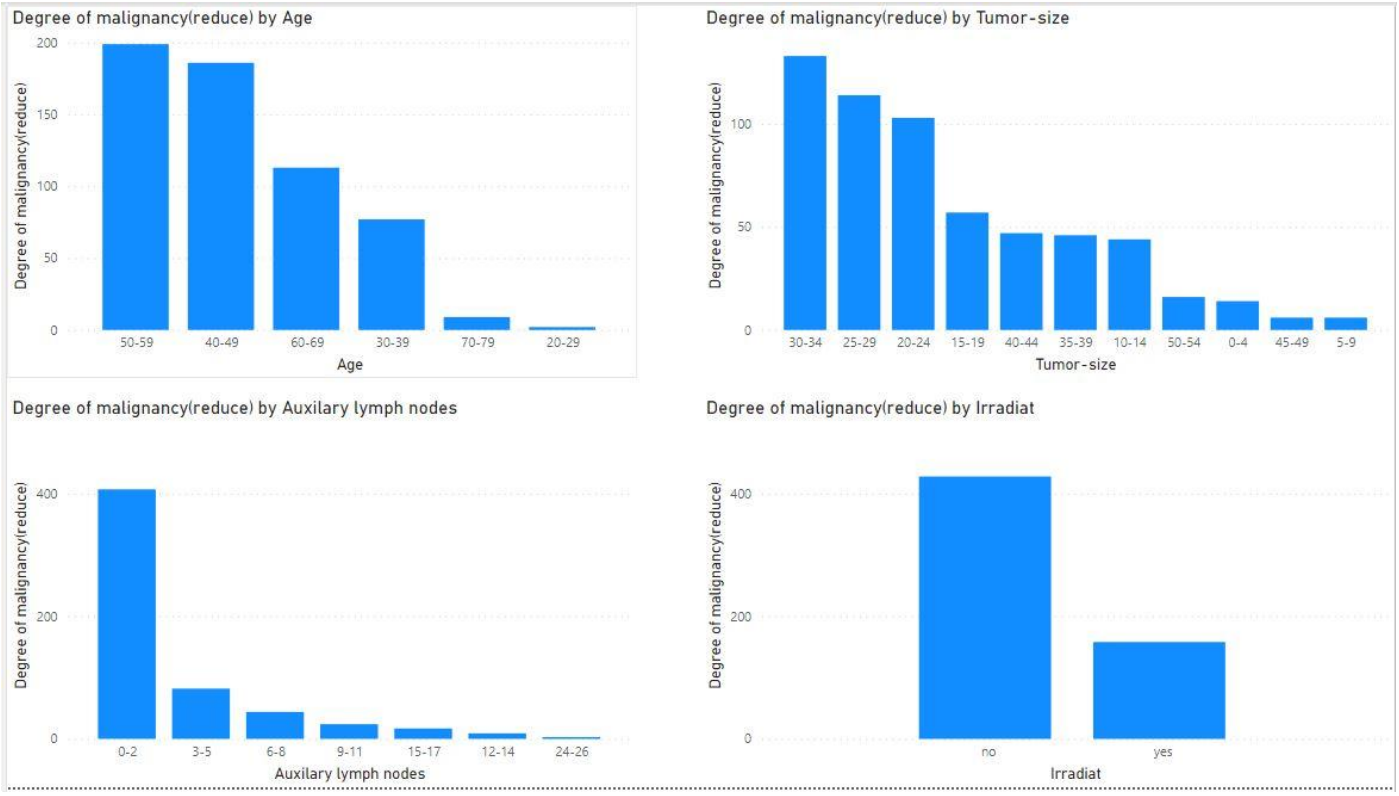


Figure 1.4

The bar charts represent the variables/factors where the spread of cancer cells is high. In other words, the dashboard shows the degree of malignance (spread of cancer cells) at different levels of age, Tumor size, exposure to radiation level (irradiate) and axillary lymph nodes. Following conclusions can be derived:

- The age group of 40-59 experiences higher risk of spread of cancer cells to other organs and tissue (degree of malignance). While the spread of disease decreases in adults with higher age group, i.e., between 60-69 and 70-79. Similar case is observed with lower age group ranging from 20-29 and 30-39.

- The spread of cancer cells based on axillary lymph nodes, the axillary lymph nodes are at different levels of the body, there are two to three nodes specific to the connection breast transporting blood, which are also called 1,2, and 3 level nodes. The bar chart represents these nodes, showing higher spread of cancer cell, i.e., 0-2 and 3-5. While the nodes more than 6 are pollable to be located in other regions of the body or close to the infected area/region.

- The size of the tumor also plays an essential role in control and spread of cancer, based on the bar graph tumor size between 20-34 have higher rate of spread of cancer cell, in comparison to ranges 15-19, 40-44, 35-39, 10-14 (medium) and 50-54, 0-4, 45-49 and 5-9 which are small.

- Next, we take a look at the exposure of cancer cells to radiation, from the bar chart we can see that the exposure of cancer cell to radiation is less, than those cells which are exposed to radiation.

The dashboard below represents the analysis done on hospital finance, cost of manufacturer products, and salary of doctors.

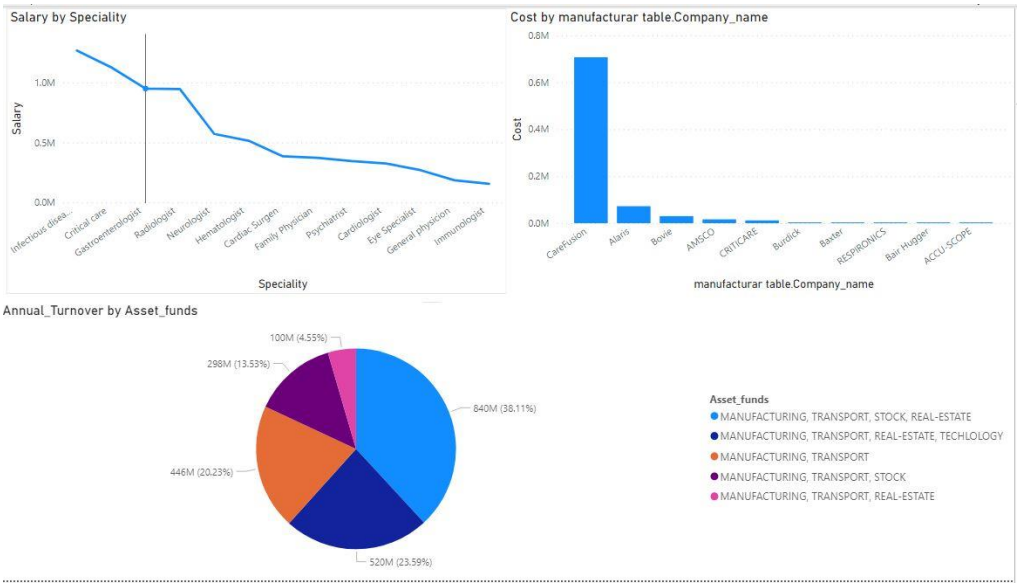


Figure 1.5

- The first graph shows the salary of the doctors based on their speciality. The line graph shows that Infectious disease specialists, critical care, gastroenterologists, radiologists, neurologists, are one of the highest paid doctors.
- Coming to the next graph, it is the cost vs manufacturer bar chart which shows the cost of the products that the manufacturer produces. Based on the costs, we can say that CareFusion, medical equipments are the most expensive.
- Next, is the pie chart based shows the annual turnover of hospital based on the investments on the asset funds.
    - Among all the investments, Manufacturing, Transport, Stock and real-estate has given the highest turnover of 38.11% i.e., 840 million. However, manufacturing, transport, real state shows least annual turnover of 100 million (4.55).

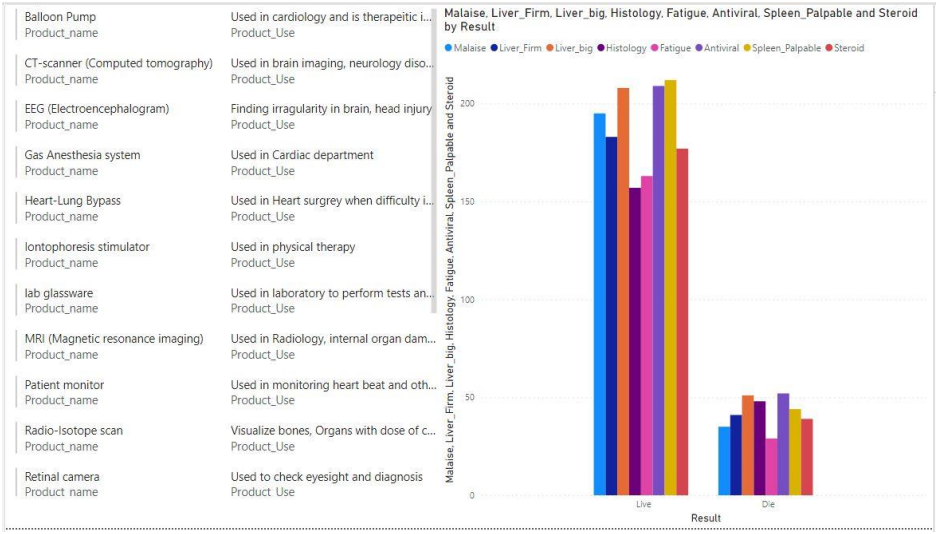


Figure 1.6

This is a report showing the live and death rate of patients and use of different equipment in different department. Based on the factors we monitor for liver condition, i.e., Malaise, Liver firmness, Liver size, Fatigue, histology, antiviral, spleen palpable levels and steroid. Patients tend to survive most often than pass away.

The dashboard below analyses, Hypertension, Stroke and Heart disease in patients:

Figure 1.7

Based on the above dashboard and graphs in it, we can make the following conclusions:

- The graph of heart disease vs gender, it observed that married male population, have higher risk of heart disease, than female population. However, unmarried people are unlikely to get affected by heart disease. One possible reason could be the food, stress levels, living in family setup. People with families tend to have proper rich food/meals, which might contain higher level of oil, butter and fats that might cause the arteries to block.

- Funnel chart on the top right, shows average glucose levels, based on the work type, with population readings on the bar. The glucose levels are higher in population with private work type, this reading reduces as we move down the graph. The most possible reason is higher stress, inactivity (IT sector work type), illness or infections.

- Another interesting observation from clustered column chart on the bottom right is, the male and female population with former or no history of smoking have experienced higher risk of heart disease.

- The fourth chart on bottom left, shows the levels of hypertension, stroke and heart disease in married and unmarried male / female population. It can be assessed that married population is more likely to face these problems, out of which hypertension is more prominent in married population.

Below figure is a report of the visualizations created in the dashboards.

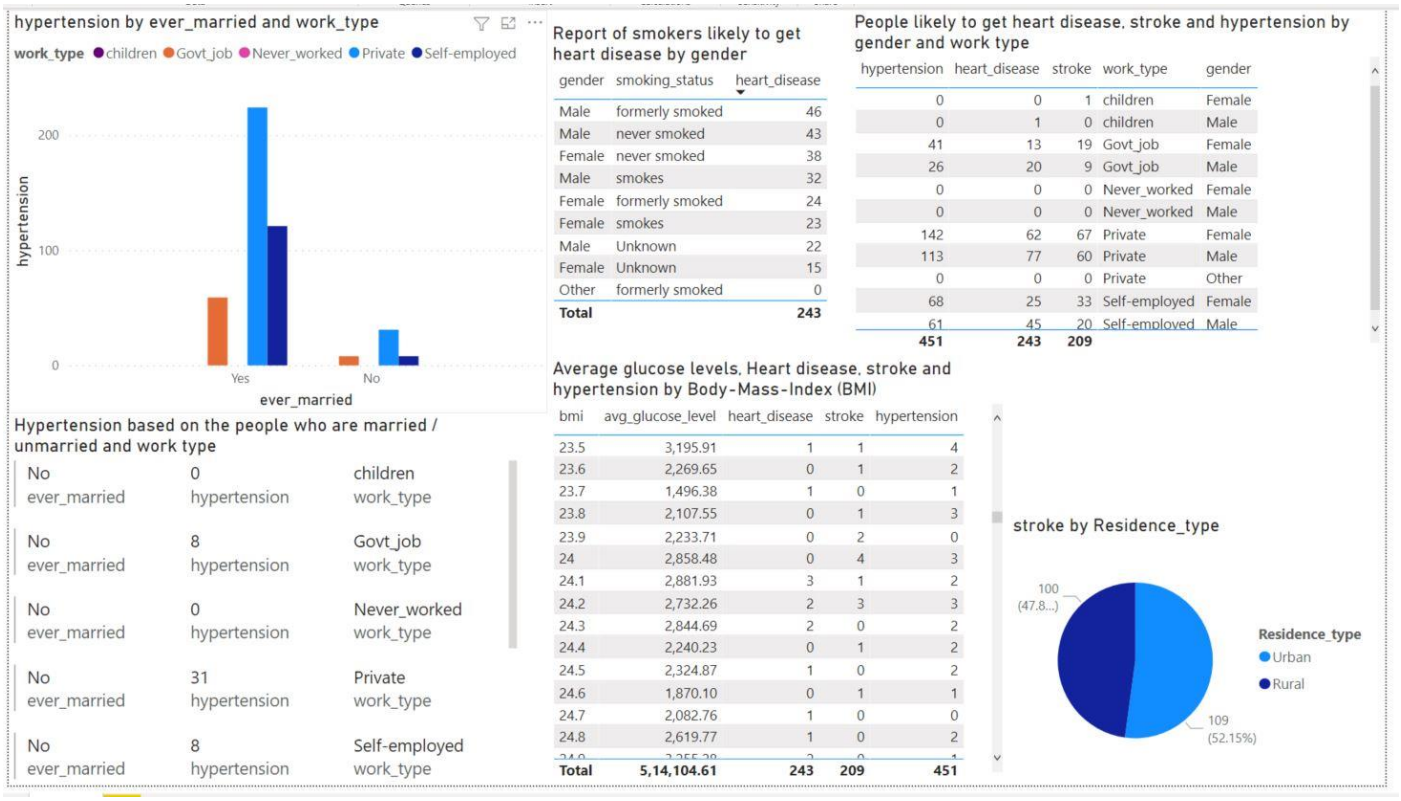


Figure 1.8

Here are a few interesting insights that we can derive from the report:

- If we evaluate the hypertension levels by work type, then population having private work type, experiences higher levels of hypertension, as compared to government and self-employed sector.
- Total number of male and female population having heart disease is 243.
- Another observation at average glucose levels base on Body Mass Index is, people with 23 and above are experiencing multiple cases of stroke and hypertension. This shows that maintain body mass index is essential to avoid stroke or hypertension.
- Stroke level by residence type, urban areas are experiencing higher cases of stroke, compared to rural area, though the difference in the number are not huge.
- Moreover, amongst the three medical conditions, i.e., stroke, heart disease and hypertension, maximum cases are observed in patients with hypertension.

KIMBALL ROAD MAP:

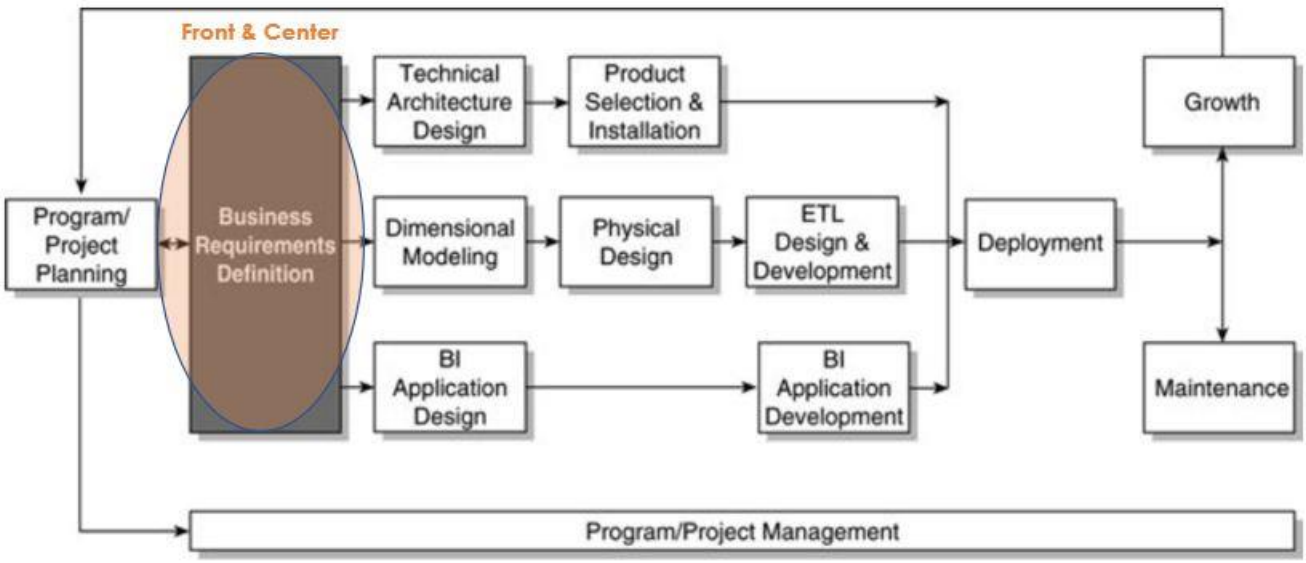


Figure 1.9

Program/Project Planning:

Since, health plays a vital role in one's wellbeing, it is essential to maintain a healthy lifestyle for people. This also makes it important for the healthcare organizations and government to leverage the healthcare services provided by them. If we turn the pages of history, the world has been affected many times by adversely infectious diseases, that have affected life, economy and growth of nations and people around the globe.

This has motivated me to come up with a project to build a data warehouse system, which allows the organizations and government institutes to be prepared for harmful and dangerous out breaks of diseases. Moving further, to accomplish this task we need data that can help us solve the objective. The business requirement here will be as follows:

- A healthcare organization (Georg healthcare) which takes an initiative to work on a project or objective to leverage the healthcare system, by analysing the symptoms of diseases, collecting data of diseases and infections currently persisting in the nature.
- Have a collaboration with organization or government to fund for research and technology, for advancement in medical research that can help improve, the process of faster diagnosis/detection of disease/medical condition.
- We need data from external and internal sources, data of patients, previous cases of diseases, information and purchase of the equipment that can accomplish the task required.

Business Requirement:

To accomplish the objective, we need to define the business process, Activities and key performance index (KPIs) and a process flow. This will help us determine the key parameters which are essential to monitor, in order to accomplish precise results in identifying and diagnosing a disease. So, we design a business process diagram, mentioning the business process, which in over case is 6 namely, (Data Collection, Symptom evaluation, Diagnosis, Prevention, Equipment evaluation and Hospital Finance). After this we generate the Key performance index for each business process, and mention the corresponding fact and dimensional tables, describing the values/components. Then we need to make a common matrix which shows, common issues or components/processes needed for different business problems.

Technical architecture/Design:

Now we need to design the technical architecture so that we can understand how the data flow in an organization will take place. The technical architecture is divided into different parts let us discuss that.

Source System data:

The data in the source system comes from different internal and external sources. The disease data, patient details, symptom evaluation, hospital finance are the types of data that will be collected from outside sources. This data is in the form of .csv, .txt and .data files, that had unstructured, structured and semi-structured data, the data did not contain columns/attributes, and had null values that needed to be removed.

The external data for Liver_disease, breast_cancer was taken from UCI_Machine Learning Repository.

The external data for healthcare data set (Patient, stroke, Heart_disease and hypertension) was taken from Kaggle

The internal data was generated for hospital finance, Doctor, Disease, Equipment, Manufacturer, Product.

ETL:

Once the data is fetched from the source, ETL operation cleans the data, structures it into the appropriate format, using normalization and other parameters. In over project the ETL is done manually given the data was not huge, but automated tools can do this task much effectively and faster. The data from external source was unstructured, there were separate files for attribute names and attribute's data. The attribute information and names of the attributes had to be merged, NULL values were removed by the mean of the column. For categorical values the rows with missing information were deleted, this reduced the actual size of the data.

After the cleaning stage, staging process is done where, the data is tables were created using SQL based on the design of the database in MySQL, where the data was supposed to be loaded once confirmed.

BI Tools:

Business intelligence tools are used to analyse the data that comes from the database and represent it as visualizations, reports using Power BI. In the BI application architecture, different tasks/operations are performed on the data like averaging specific insightful columns/attributes, comparing survive to death ratio, etc. These insights allow the team to understand the pattern of the virus or the cause of a medical condition.

User Interaction:

The visualizations generated by the BI team, helps the users to draw insight and understand useful information about the objectives the business user is trying to accomplish. Since, the predictions and analytics are represented in the form of dashboards and reports, the details of the analytics can be easily conveyed and communicated to the business user via dashboards and reports/visualizations or visual stats. This includes bar charts, histograms, area plots, maps, and other visualizations. We have used PowerBI and made a few dashboards that convey/ visualize the chances and possible causes of hypertension, stroke, heart disease and liver_disease in the patients.

Dimensional modeling:

Dimensional model is built based on the design of the database model, which is an entity relationship diagram of fact and dimension tables. Fact tables contain measure values or values that can be measured, the fact tables are connected with dimension table by a foreign key, a fact table is the primary table in a dimensional model.

A dimensional table contains dimensions of a fact table, and contain descriptive information about the facts which are in the form of attributes. Generally, a dimensional table contains values which are non-numeric and categorical, but it can contain numeric values which do not change repeatedly or frequently.

Based on Kimball approach, first based on the business process we need to identify the granularity, or the level of detailed description of an event or object. It is like breaking a process into small components and listing the required attributes in to the table. Once we are done with that, first we generate the dimensional table and then generate the fact table, which is the main component of the database design and dimensional model.

The fact table can be represented as a star schema which is most effective because it offers performance enhancements, where else snowflake schema can have one dimensional model connected to another, it is a complex design and requires multiple joints in order to access information. Which in the case of star schema is not the problem, and hence the star schema is preferred design which is implemented in this project.

We have Equipment, hospital, report and disease fact tables with their respective dimensional tables, to generate an effective database design we can aggregate the fact tables as one tables namely Hospital_Fact and connect it with the dimensional tables making it a star schema as shown below.

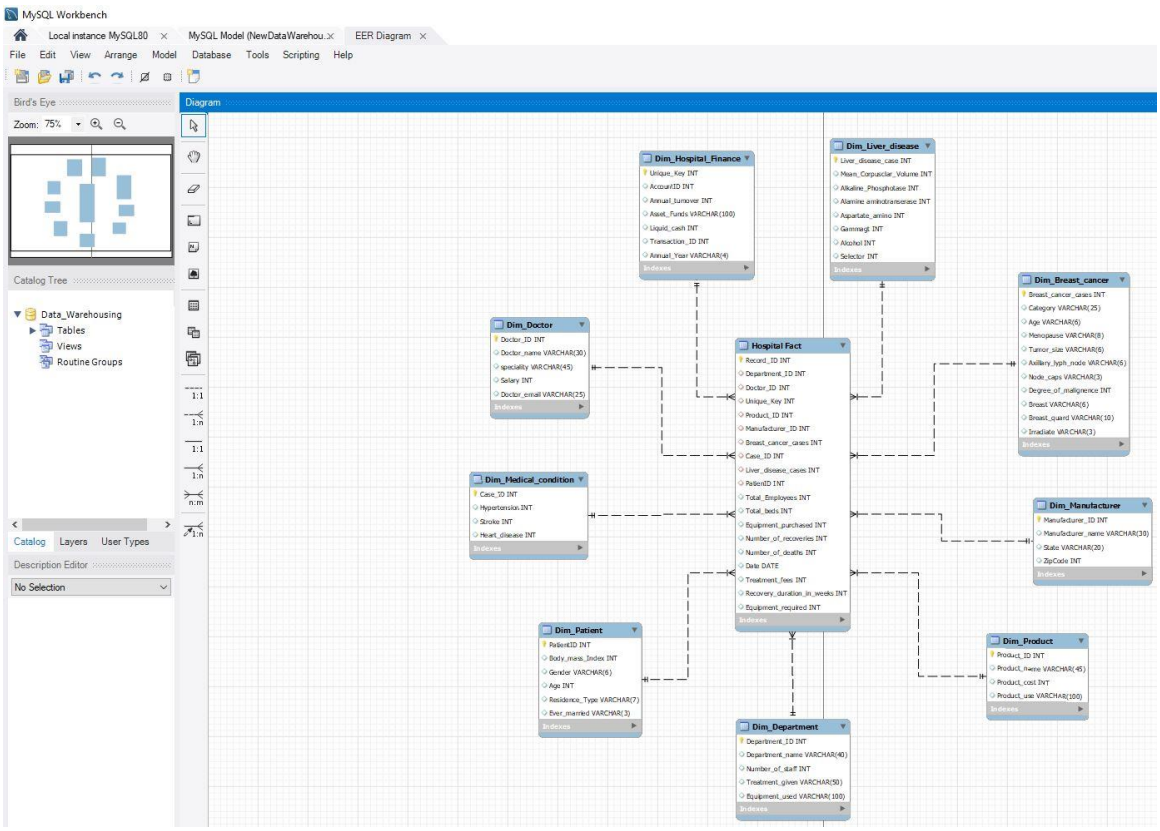


Figure 2.0

This star schema of hospital disease diagnosis has one fact table in the centre with 9 dimensional tables. Namely Patient, Department, doctor, Brease cancer, Liver disease, medical condition, product, manufacturer and hospital finance.

Physical Design:

The physical design consists of the database consists of removal of null values, standardization of the information, staging, primary and foreign key relationship. Here we specify certain standards with the help of which, we can reduce confusion, mismatching and repeatability in the relational database model. In the current project standardization is used on the contact information of doctor's email Id, on Product name and Manufacturer_ID, and while accessing hospital finance information.

The doctors email contains character standardization with first two letters of first and last name followed by the hospital name's, @ and .org (organizational email).

Manufacturer information contains numeric standardized rules, and product name consists of character standard rule, with first three letters taken from company name followed by the product name. This ensures that specific product is sold by a specific manufacturer and it is easy to identify product information.

ETL Design and Development:

ETL design and development phase consists of 4 major components, Extracting, Cleaning and Confirming, Delivering and Managing. This is an essential process that helps to modify and shape the data before embedding it into the database.

Extracting:

The data extraction was done from external sources and data was also generated internally. External data was taken from UCI Machine Learning repository for Breast cancer and Liver disease symptom evaluation and disease diagnosis and health care data was taken from Kaggle repository for general medical condition and heart disease problems.

Cleaning and Conforming:

The data was in the form of .csv, .data and .text file, apart from the csv file the data was unstructured and it had to be arranged. First the data file with information of attributes is modified by adding the column attributes in the file. Later, the number of missing values are identified in the data, all this is done in python Jupyter notebook and MySQL using SQL query. The null values are removed by taking the mean for numeric values and deleting the rows in case of categorical missing values. The data is saved once cleaned.

Delivering:

The data is then loaded into the database, in structured format. Since the data keeps changing in specific duration of time, for instance monthly, then we can store the data in the form of refresh, this will keep the data updating and refreshing as any new modifications are done. Tools like PowerBI provide this functionality where we have an option to save the data as refresh.

Managing:

Reviewing and revising the data, to see if there are any lags, if there are lags we need to perform checks on the relation ship between the tables, check for fact tables, dimension tables, OLAP processing structure, for proper functioning.

Business Intelligence Application:

In business intelligence application, task like analytics and presentation of the data takes place, here the data from the data warehouse is used by the business intelligence team, analysts, data scientists, of the organization, to built meaningful visualizations for the business user to understand the insightful information gathered by the processing of data.

The insights from the data are derived by using many processes, like data mining (clustering, segmentation, classification techniques), data visualization, creating graphs, bar charts, histograms, tree maps, bubble plots. Relationship graphs like scatter plots and more. These insights help the organization to represent the relationship between the target attributes to the business user on which the project is focused on.

Different dashboards and reports are created containing, pie chart, bar charts, histogram and line charts, representing the factors responsible for a medical condition in patients. This also includes the cases of breast cancer, liver_disease and hepatitis conditions.

Business Application Development:

In this process we use BI development tools/ methods to develop a optimal and scalable BI solution to over problem. To accomplish the development process we should use dashboards, reports and work with business owners to ensure high quality and performance. In over case we can communicate with the government institutes and health organizations by presenting the dashboard and visual analytical designs, to mitigate the short comings and loopholes in healthcare system, which are retrieved from analytics and BI team.

This helps the organization and business users to brainstorm on any other development tools which can leverage and clean the system blocks.

Deployment and maintenance:

Deploying a system is essential to understand whether the sub systems, embedded in a system work in coordination with the integrated system components, and is efficient in achieving/fulfilling the project and business user requirements/objective.

There are different parameters and guidelines an organization can form to deploy a model; these guidelines are composition of distinct methods or processes used to deploy a project model or system.

For healthcare and disease diagnosis following deployment and maintenance processes can be performed:

- First, we need to monitor the checklist and guidelines that the business users (Government, researchers, hospital stakeholders) have provided the organisation.

- Need for index tuning (for faster query processing and fetching specific data.), performance tuning (improve performance allows for better results from the system), system backup (essential when technical/system/software updates or services take place in an organization).

- Testing, one of the most essential components in deploying a system in an organization. To deploy a complete system, Unit testing, User Acceptance Test, Trace Matrix and Software Development Life Cycle models can be used.
    - Unit testing is performed by a database programmer or by programming, the most important aspect of disease diagnosis, is to identify the symptoms that are inclined in distinguishing one disease. This part of the system needs to be accurate and provide precision, so that the patient can be given, proper treatment resulting in faster recovery.

- User Acceptance Testing, is important because the business user need to be satisfied by the results / output that the system delivers. If the results are not up to the users' expectations, then there is a possibility to re-evaluate the entire process in the system, since all the systems are interrelated to each other. To avoid this the business user is involved in the process of user acceptance testing, where the user is allowed to get exposed to the system operation. They are asked if it meets their demands, if yes then further operations / modifications are done in the system.

The run team later maintains the organizations operations and keeps them operation ready.

Growth:

The healthcare system is large and contains many components that are essential for consideration while growth of an organization.

- Space/ infrastructure, setting up a healthcare facility requires infrastructure, space, designated buildings for specific departments. For example, laboratory, radiology, orthopaedic department etc.
- People to manage and maintain resources (equipment) in infrastructure. We need skilled workers to operate the equipments and assist doctors while performing surgery in OT.
- Skilled workers, doctor's tech support, management, stakeholders.
- Research laboratory and access to information on distinct and diverse pathogens, bacteria, virus.
- Collaboration and tie-ups with local and state level healthcare facilities/organizations to develop and monitor the needs in healthcare facilities, research data and equipment, promote collaborative work share information and work together.
- Funding from government or financial giants to foster advance research and invention of effective and safe drugs.
- Safety and regulations guidelines and facilities to perform medical tests based on, on-going research and inventions.
- Collaboration with Technology providers to maintain and manage health data, get fruitful insights to help and understand the depth of research components. This is where business application and data driven analytical processes come in. where IT companies help the researchers or scientists and government with tech support to get more insight towards, the organization objective and disease diagnosis.
    - The technical team collects the information and migrates it to data warehouse or data lakes for easy access and faster retrieval of symptom similarity to an unknown disease or pathogen.
    - Data Governance and data management for metadata, monitoring organizational data come into consideration here.
- Finally, work culture also impacts the growth of an organization. We want to develop a work culture where a single person is not responsible to perform a specific duty or type of work. We want to train and give chance to assisting doctors or healthcare worker to take up responsibility. This can be done under the observation of an experience person, so that whenever necessary or emergency situation come up. The cases can be distributed amongst doctors establishing an efficient workforce.