# Solution - *Explain your solution here in a step by step manner.*

* At first, upload all the data in AWS s3 bucket inside a folder input-data.
* And use Data Bricks to create multiple notebooks, where all the data will be read by making data frames like df\_claims, df\_disease,df\_group as per the given datasets.
* Now will use different SQL and spark functions to clean the data and join it to create a dimensional model using snowflake model, like fact and dimensions.
* After cleaning and joining the datasets the tables will created and transferred to redshift tables.

# Use Cases - *List down all the use cases on which this solution will be applicable.*

The use cases that are going to resulted from the redshift table using SQL queries.

* Which disease has a maximum number of claims.
* Find those Subscribers having age less than 30 and they subscribe any subgroup
* Find out which group has maximum subgroups.
* Find out hospital which serve most number of patients
* Find out which subgroups subscribe most number of times
* Find out total number of claims which were rejected
* From where most claims are coming (city)
* Which groups of policies subscriber subscribe mostly Government or private
* Average monthly premium subscriber pay to insurance company.
* Find out Which group is most profitable
* List all the patients below age of 18 who admit for cancer
* List patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.
* List female patients over the age of 40 that have undergone knee surgery in the past year

1. Database Design - List down all possible db(Redshift) tables here

## Tables Metadata Info with Pk/FK relationship –

|  |  |
| --- | --- |
| **Claims** | **Disease** |
| Claim\_or\_rejected | SubGrpID(FK) |
| SUB\_ID(FK) | Disease\_ID |
| claim\_amount | Disease\_name(PK) |
| claim\_date |  |
| claim\_id(PK) |  |
| claim\_type |  |
| disease\_name(FK) |  |
| patient\_id(FK) |  |

|  |  |
| --- | --- |
| **Group** | **Grpsubgrp** |
| Country | SubGrp\_ID(FK) |
| Premium\_written | Grp\_Id(FK) |
| Zipcode |  |
| Grp\_Id(PK) |  |
| Grp\_Name |  |
| Grp\_Type |  |
| City |  |
| year |  |

|  |  |
| --- | --- |
| **Patient\_record** | **Sub Group** |
| Patient\_id(PK) | SubGrp\_id(PK) |
| Patient\_name | SubGrp\_Name |
| Patient\_gender | Monthly Premium |
| Patient\_birth\_date |  |
| Patient\_phone |  |
| Disease\_name(FK) |  |
| City |  |
| Hospital\_id(FK) |  |

|  |
| --- |
| **Subscriber** |
| Sub\_id(PK) |
| First\_name |
| Last\_name |
| Street |
| Birth\_date |
| Gender |
| Phone |
| Country |
| City |
| Zip Code |
| Subgrp\_id(FK) |
| Elig\_ind |
| eff\_date |
| term\_date |

* Red Shift Table according to use cases

1. Which disease has a maximum number of claims?

|  |
| --- |
| **DiseaseMaxClaim** |
| Claim\_id |
| Disease\_id |
| Disease\_name |
| Claim\_amount |

1. Find those Subscribers having age less than 30 and they subscribe any subgroup

|  |
| --- |
| **Subscriber** |
| Sub\_Id |
| SubGrp\_Id |
| First\_name |
| Last\_name |
| Age |

1. Find out which group has maximum subgroups.

|  |
| --- |
| **Grpmaxsubgrp** |
| Grp\_Id |
| SubGrp\_id |
| Grp\_name |
| SubGrp\_name |

## ER diagram - *Optional*

# Technologies and Platforms to be used in this solution -*List down list of technologies like spark, aws and databricks etc.*

1. AWS s3
2. AWS RedShift
3. Databricks
4. AWS EMR Studio
5. Pyspark
6. Jira
7. GitHub

## 