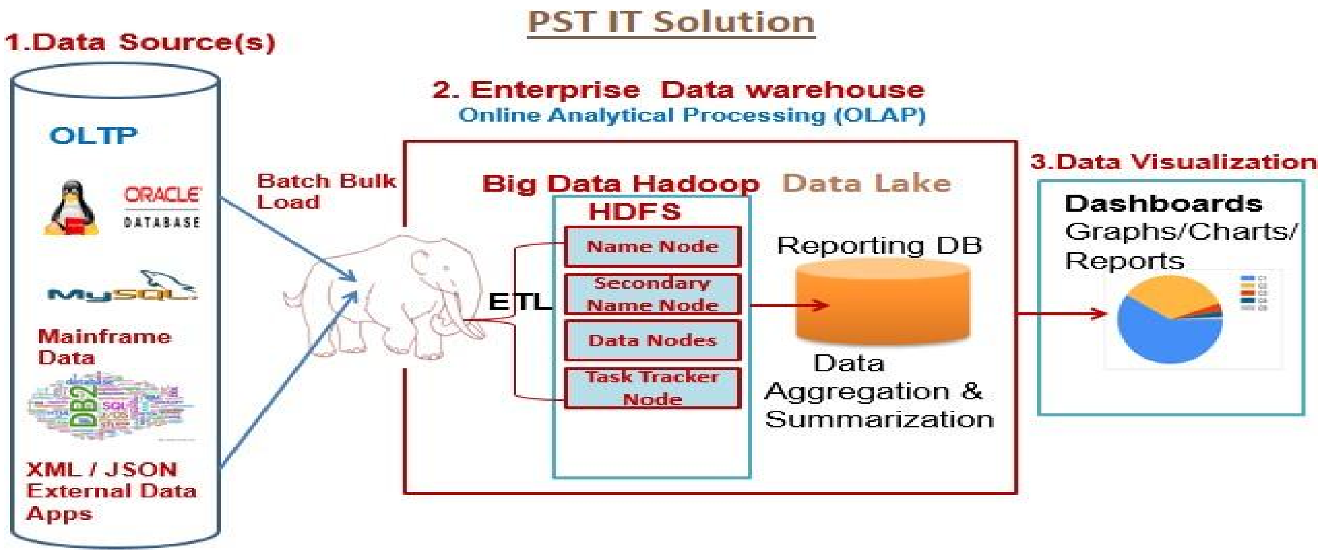
HEALTH CARE SYSTEM ANALYSIS

THE ASSURANCE OF EXCELLENCE
PST
IT SOLUTIONS

Contents

THE ASSURANCE OF EXCELLENCE
PST
IT SOLUTIONS

Data Migration Process :-



I have around 3 years of IT experience. 2 years of working in Hadoop Big Data from work In Company Name(currently working). I am working for **client name** and the project was Data Ingestion into Data Lake. So **client name** as Hadoop vendor they areusing Cloudera that is own premises. As of now the Hadoop cluster is not on cloud so the cluster is on own premises, So we have migrated data from n number of sources of**client-name** into data lake.

Client-name having 100s of data source and multiple datasets are there, so thatwe build a data ingestion framework. And that data ingestion framework support 3 ingestion pattern which are

1. Data Base Ingestion
2. File Based Ingestion
3. CDC(Change Data Capture) or we can say Incremental data load

My role was a Developer, and also involved in the 1st part of the project this is called data acquisition. We have to meet with the client and requester who needs to inject their data so we have to provide a solution based on that source system.

Like :-

1. What type of source system is this.
2. What type of data they are having and
3. How much data is there

So based on these we are advising that which type of data ingestion pattern we have to use.

So mainly

1. In data base ingestion we are using the sqoop.
2. In the file based ingestion we are using Hadoop commands and shell scripting.
3. And 3rd one is CDC(Changed Data Capture) In that we are using one 3rd party tool i.e. Attunity.

I. So in database ingestion we are like directly connecting the source databaseand we are importing the data into data lake.

II. In the file base ingestion or in some sources, they are not providing the readaccess of their databases for that they are extracting a data into file and theyare sending a file into our edge node in the landing area.

So there are two ways we set up this framework or environment to transfer a file.

1. One is SFTP and
2. Connect Direct Gateways.

So in a file based there are one more category like they can send

1. Fixed length file
2. Delimited file
3. JSON or XML file

So once the data is in our landing zone, we are team is like we are preparing a code for that and then we are creating the DDL and configuration and we are creating a table. Ultimately, we are ingesting all data into a hive table only.

So this is all the phases like we have to go through the UAT development and then inproduction.

The fist production activity we have to do manually like the first ingestion we have to perform manually and then we have to setup a **control-M** based on that ingestion frequency agreed on this **d**ata acquisition part with the source team.

So in data lake, This is the central repository and there nobody is allow to transform, update or change the data in data lake.

So for consumption purpose, we are propagating the data into some other cluster,those are also physical or separate cluster for business wise and as it is like we are propagating the data in the consumer cluster.

Business Challenge / Requirement

A Health Care insurance company is facing challenges in enhancing its revenue and understanding the customers so it wants to take help of Big Data Ecosystem to analyze the Competitors company data received from varieties of sources, namely through scrapping and third-party sources. This analysis will help them to track the behavior, condition of customers so that to customize offers for them to buy insurance policies and also calculate royalties to those customers who buy policies in past, this in turn will enhance their revenues.

1. The goal of the project

The goal of the project is to create data pipelines for the Health Care insurance company which will make the company make appropriate business strategies to enhance their revenue by analyzing customers behaviors and send offers and royalties to customers respectively.

2. Data Flow Architecture / Process Flow

1. Multiple Data base sources and Edge node receives data files in form of json ,csv and XML. These files are coming from the sources based on user interaction methodology.

2. The files data is validated, enriched and processed before loading into HDFS System.

3. After validated the files data, we are creating the data model for HDFS in HIVE so that we can storethe files data into the HDFS.

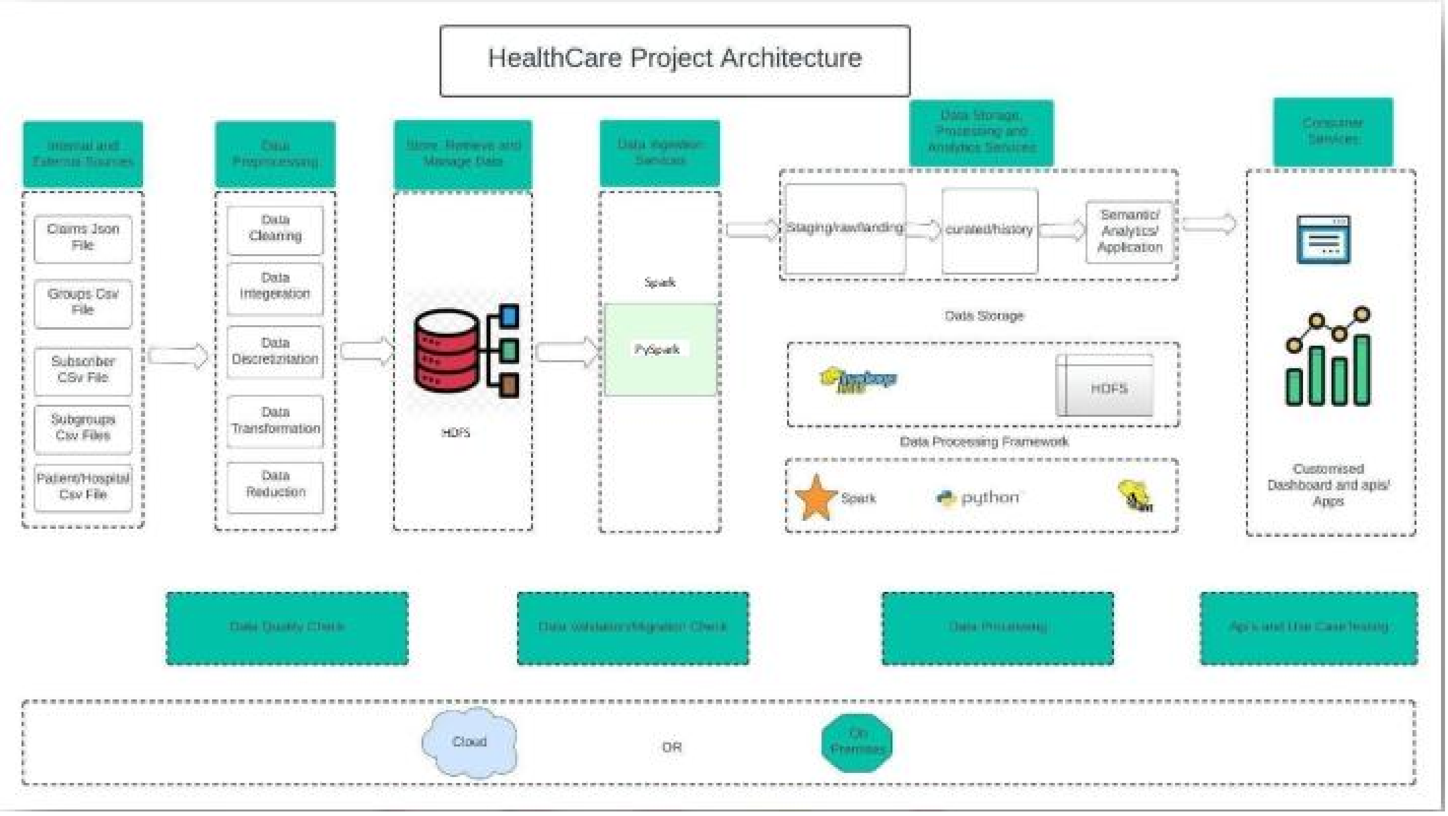
4. After storing the data into the HDFS, we now transform according to our business requirements.

5. Finally, data landed to again HDFS needs to be analyzed by some analytical queries.

6. After analytics queries, we test the result and use the result to enhance the company revenues.

A schematic flow of operations with the best suited components is shown below

THE ASSURANCE OF EXCELLENCE
PST
IT SOLUTIONS



3. Dataset Explanation & Schema

Data coming from third-party sources reside in local directory and has csv and json format.

Fields present in the data files and tables-

Data files contain the below fields.

Column Name/Field Name Column Description/Field Description

Json File Fields

- CLAIM_ID
- PATIENT_ID
- DISEASE_NAME
- SUB_ID
- CLAIM_DATE
- CLAIM_TYPE
- CLAIM_AMOUNT
- CLAIMED_OR_REJECTED

CSV File 1 Fields (Patient.csv)

- PATIENT_ID
- PATIENT_NAME
- PATIENT _GENDER
- PATIENT _BIRTHDATE

THE ASSURANCE OF EXCELLENCE

- PATIENT _PHONE
- HOSPITAL_ID
- DISEASE _ NAME
- CITY

CSV File 2 Fields (Subscriber.csv)

- SUB_ID
- FIRST_NAME
- LAST_NAME
- STREET
- BIRTH_DATE
- GENDER
- PHONE_NO
- COUNTRY
- CITY
- ZIP_CODE
- SUBGRP_ID
- ELIG_IND
- E_DATE
- T_DATE

CSV File 3 Fields (Group.csv)

- GRP_ID
- GRP_NAME
- PREMIUM_WRITTEN
- GRP_TYPE
- PIN_CODE
- CITY
- COUNTRY
- ESTABLISHMENT_YEAR

CSV File 4 Fields (disease.csv)

- SUBGRP_ID
- DISEASE_NAME
- DISEASE_ID

CSV File 5 Fields (subgroup.csv)

- SUBGRP_ID
- SUBGRP_NAME
- GRP_ID

CSV File 6 Fields (hospital.csv)

- HOSPITAL_ID
- HOSPITAL _ NAME

CITY
STATE
COUNTRY

THE ASSURANCE OF EXCELLENCE
PST

4. Problem Statements / Tasks

5.1 Problem 1- Data Pre-processing, Enrichment and Load into Database

- Parse and Infer schema of the given xml and csv formats data is ingested.
- We are expected to do general data cleaning steps like empty string replacements with actual NULL, data type checks (including date format) and corrections/ rejections, file name checks, empty file checks, malformed record checks and rejection etc.

Learners must apply below rules for data enrichment process:

Validate the data from the input file and load only valid records into the target table according to the constraints mentioned in the target table.

Load only the members who are currently effective. (i.e.) SYSDATE BETWEEN EFFT_DT AND TERM_DT

Reject records if the Subscriber_Id has less than 9 characters.

Populate leading zeroes in the fields GROUP_ID and SUBGRP_ID while populating data into the Target table.

Also validate the Group Id and Subgrp_Id against the Subgrp table and load only matching data into the target table

Schema Design for SQL Database

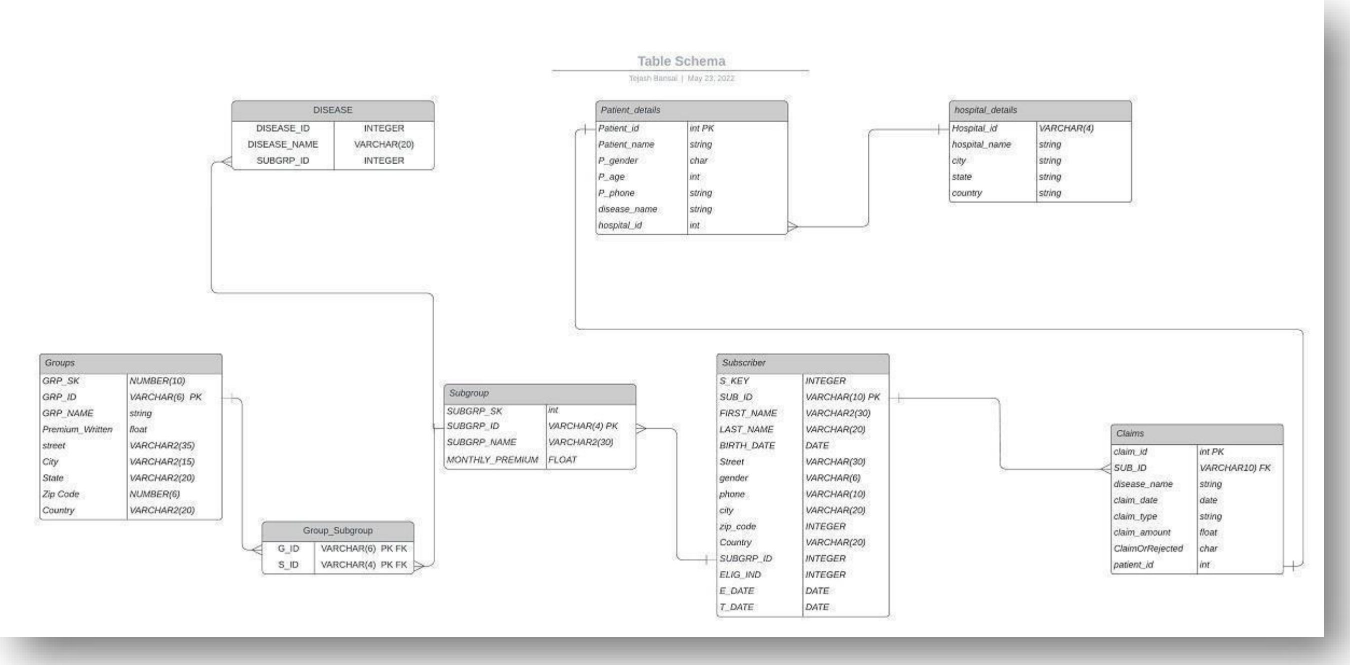


Table Schema

Tejash Bansal | May 23, 2022

DISEASE	
DISEASE_ID	INTEGER
DISEASE_NAME	VARCHAR(20)
SUBGRP_ID	INTEGER

Patient_details	
Patient_id	int PK
Patient_name	string
P_gender	char
P_age	int
P_phone	string
disease_name	string
hospital_id	int

hospital_details	
Hospital_id	VARCHAR(4)
hospital_name	string
city	string
state	string
country	string

Groups	
GRP_SK	NUMBER(10)
GRP_ID	VARCHAR(6) PK
GRP_NAME	string
Premium_Written	float
street	VARCHAR2(35)
City	VARCHAR2(15)
State	VARCHAR2(20)
Zip Code	NUMBER(6)
Country	VARCHAR2(20)

Group_Subgroup	
G_ID	VARCHAR(6) PK FK
S_ID	VARCHAR(4) PK FK

Subgroup	
SUBGRP_SK	int
SUBGRP_ID	VARCHAR(4) PK
SUBGRP_NAME	VARCHAR2(30)
MONTHLY_PREMIUM	FLOAT

Subscriber	
S_KEY	INTEGER
SUB_ID	VARCHAR(10) PK
FIRST_NAME	VARCHAR2(30)
LAST_NAME	VARCHAR(20)
BIRTH_DATE	DATE
Street	VARCHAR(30)
gender	VARCHAR(6)
phone	VARCHAR(10)
city	VARCHAR(20)
zip_code	INTEGER
Country	VARCHAR(20)
SUBGRP_ID	INTEGER
ELIG_IND	INTEGER
E_DATE	DATE
T_DATE	DATE

Claims	
claim_id	int PK
SUB_ID	VARCHAR10) FK
disease_name	string
claim_date	date
claim_type	string
claim_amount	float
ClaimOrRejected	char
patient_id	int

5.2 Problem 2 - Data Analysis (Spark/Hive)

Once we have made the data ready for analysis, we have to perform below analysis on a batch basis.

1. Find those Subscribers having age less than 30 and they subscribe any subgroup. The output can be in form of a file with columns.

COUNT_OF_SUBSCRIBER

2. Which groups of policies subscriber subscribe mostly Government or private. The output can be in form of a file with columns.

GRP_TYPE,
COUNT(GRP_ID)

3. List female patients over the age of 40 that have undergone knee surgery in the past year. The output can be in form of a file with columns.

PATIENT_NAME

4. Give the Most Profitable subgroup which subscribe the greatest number of times. The output can be in form of a file with columns.

SUBGRP_NAME,
COUNT

5. Give out which groups has maximum subgroups (Policies Groups). The output can be in form of a file with columns.

G_ID,
SUBGRP _COUNT

6. Give the result from where most of the claims are coming (city). The output can be in form of a file with columns.

CITY,
MAX_CLAIM

7. List all the patients whose age is below 18 and who admit for cancer in the hospital. The output can be in form of a file with columns.

PATIENT _ID,
PATIENT _NAME,
AGE

8. List patients who have cashless insura0nce and have total charges greater than or equal for Rs. 50,000. The output can be in form of a file with columns.

PATIENT _ NAME,
PATIENT _GENDER,
PATIENT _BIRTH_DATE

9. Find out total number of claims which were rejected by the groups (insurance companies). The output can be in form of a file with columns.

CLAIM_OR_REJECTED,
COUNT_CLAIM_ID)

10. Give out which disease having maximum number of claims. The output can be in form of a file with columns.

DISEASE_NAME,
COUNT_CLAIMS

Store the above analyzed results as a separate dataset in HDFS.

Approach to Solve

Below steps can be taken to start solving the project problem statements:

- Start by generating Raw Data files in Gateway node location
- Problem 1: Write code to clean & transform data according to the use cases and saved inside the /Processed Data/files folder. After that perform some EDA on top of the cleaned data. Write code and run to take data from /processed data /files and stores all the files in the SQL database using the python and MySQL connector.
- After that we have to write some Sqoop scripts for importing the data from RDBMS System to the HDFS directory /user/hive/warehouse/HEALTHCARE.DB/files
- Write code and run to take data from /user/hive/warehouse/HEALTHCARE.DB/files and solve Problem 2 in a PYSPARK Batch
- Write code and run to take data from '/spark output/files' and perform some visualization on that output files.
- At the end we test all use cases according to the business.

THE ASSURANCE OF EXCELLENCE PST IT SOLUTIONS

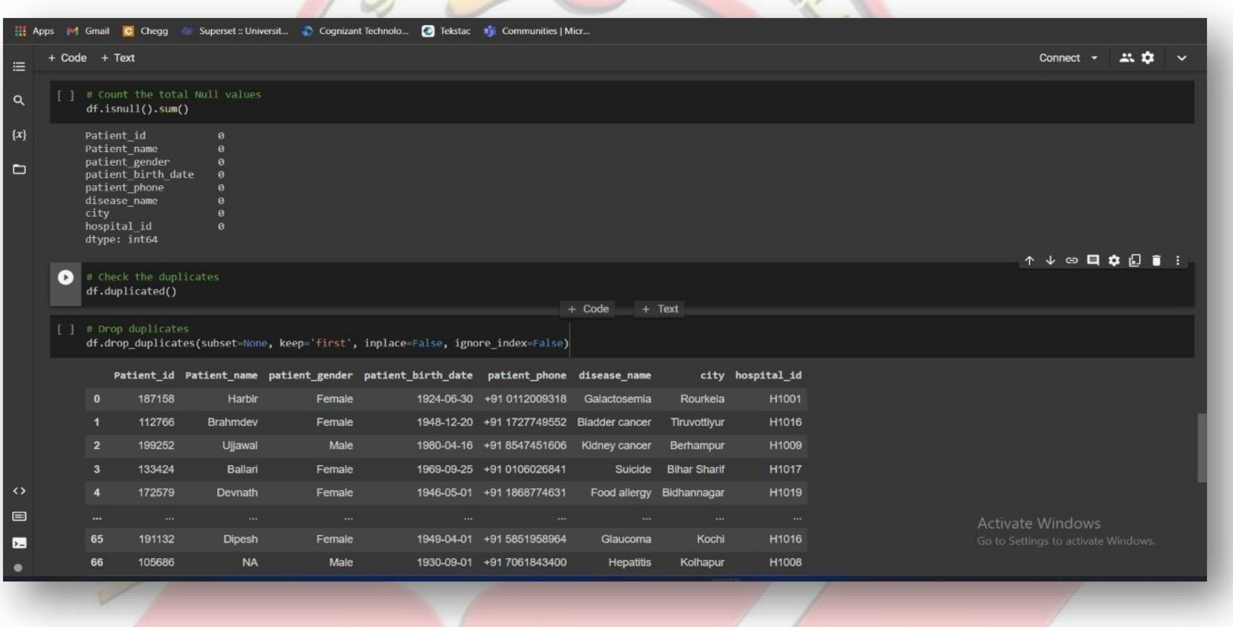
6. Coding/Code Templates:

6.1 Data Processing

6.1.1 Conversion of raw data to processed data:

For each raw file we have checked null values, duplicate values and other parameters and then converted into

processed dataset. here are some samples of codes.



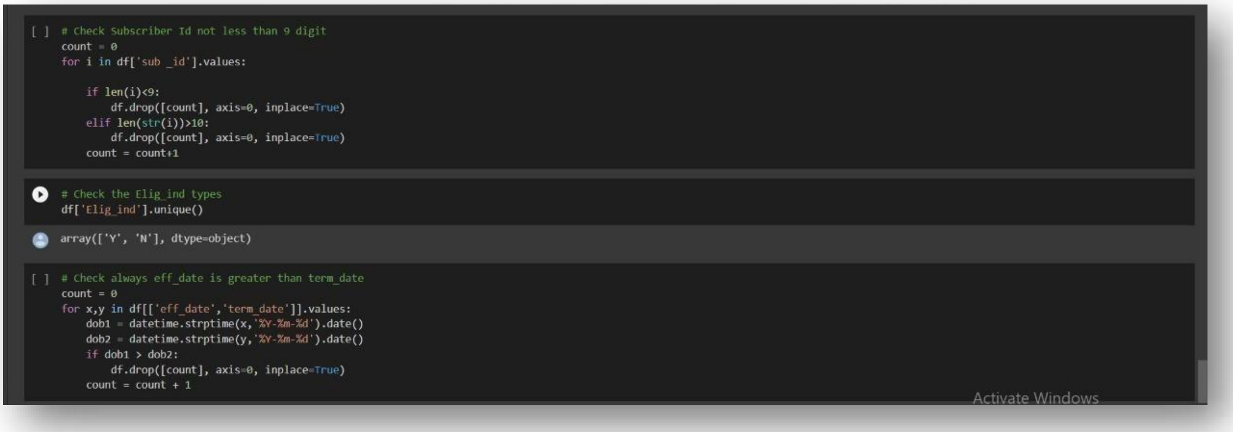
```
[ ] # Count the total Null values
    df.isnull().sum()

    Patient_id           0
    Patient_name         0
    patient_gender       0
    patient_birth_date   0
    patient_phone        0
    disease_name         0
    city                 0
    hospital_id          0
    dtype: int64
```

```
# Check the duplicates
df.duplicated()
```

```
[ ] # Drop duplicates
    df.drop_duplicates(subset=None, keep='first', inplace=False, ignore_index=False)
```

	Patient_id	Patient_name	patient_gender	patient_birth_date	patient_phone	disease_name	city	hospital_id
0	187158	Harbir	Female	1924-06-30	+91 0112009318	Galactosemia	Rourkela	H1001
1	112766	Brahmdev	Female	1948-12-20	+91 1727749552	Bladder cancer	Tiruvottiyur	H1016
2	199252	Ujjawal	Male	1980-04-16	+91 8547451606	Kidney cancer	Berhampur	H1009
3	133424	Ballari	Female	1969-09-25	+91 0106026841	Suicide	Bihar Sharif	H1017
4	172579	Devnath	Female	1946-05-01	+91 1868774631	Food allergy	Bidhannagar	H1019
...	...	...	...	...	...	...	...	...
65	191132	Dipesh	Female	1949-04-01	+91 5851958964	Glaucoma	Kochi	H1016
66	105686	NA	Male	1930-09-01	+91 7061843400	Hepatitis	Kolhapur	H1008

Here we are checking for SUB_ID if is it length of 9 or not



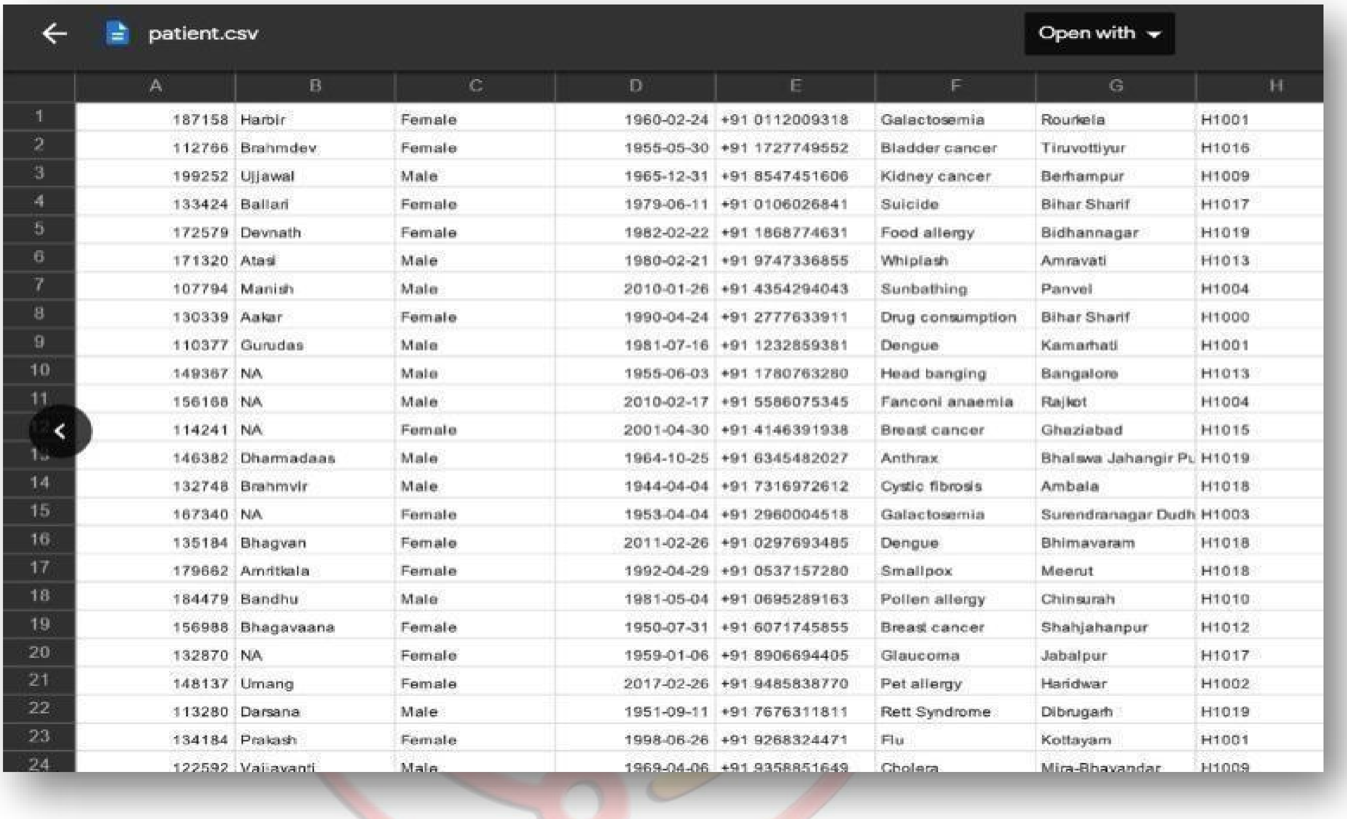
```
[ ] # Check Subscriber Id not less than 9 digit
    count = 0
    for i in df['sub _id'].values:

        if len(i)<9:
            df.drop([count], axis=0, inplace=True)
        elif len(str(i))>10:
            df.drop([count], axis=0, inplace=True)
        count = count+1
```

```
# Check the Elig_ind types
df['Elig_ind'].unique()

array(['Y', 'N'], dtype=object)
```

```
[ ] # Check always eff_date is greater than term_date
    count = 0
    for x,y in df[['eff_date','term_date']].values:
        dob1 = datetime.strptime(x,'%Y-%m-%d').date()
        dob2 = datetime.strptime(y,'%Y-%m-%d').date()
        if dob1 > dob2:
            df.drop([count], axis=0, inplace=True)
        count = count + 1
```

6.1.2 Processed Dataset

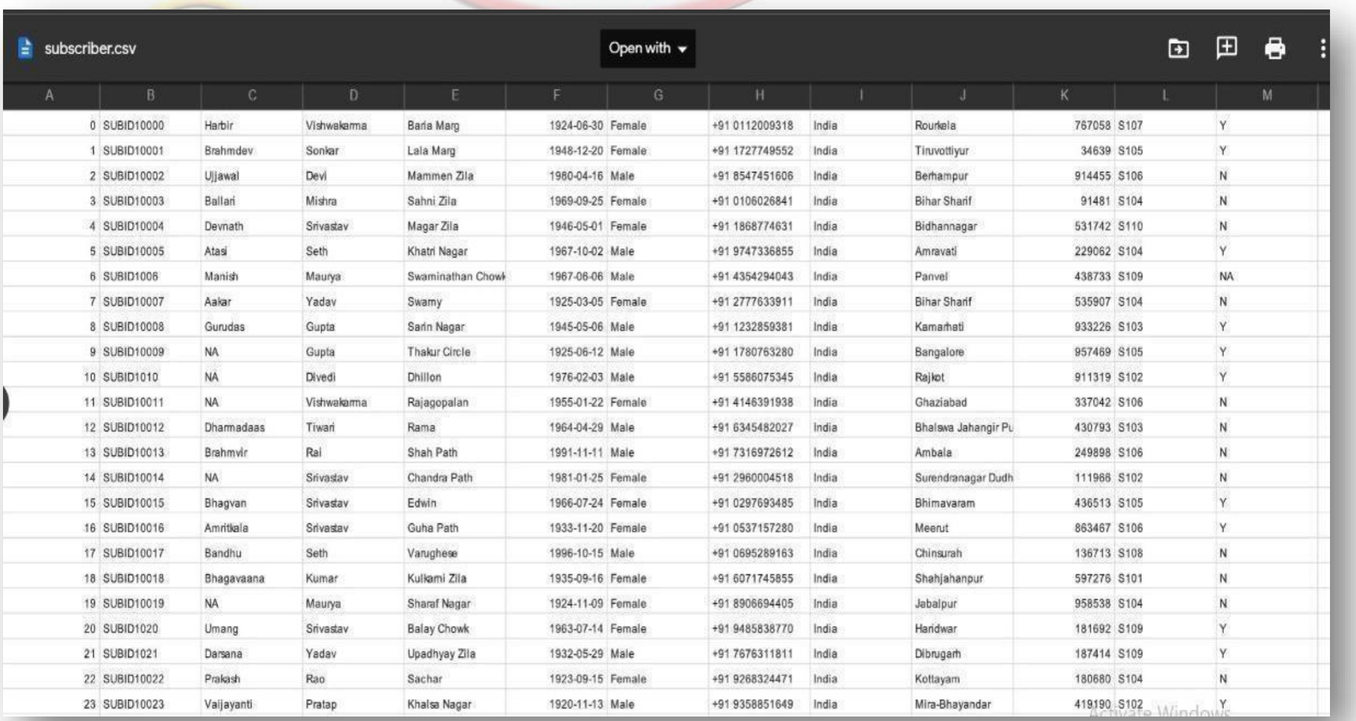
Some snippets of processed dataset which is further used to create RDBMS.



patient.csv

Open with

	A	B	C	D	E	F	G	H
1	187158	Harbir	Female	1960-02-24	+91 0112009318	Galactosemia	Rourkela	H1001
2	112766	Brahmdev	Female	1955-05-30	+91 1727749552	Bladder cancer	Tiruvottiyur	H1016
3	199252	Ujjawal	Male	1965-12-31	+91 8547451606	Kidney cancer	Berhampur	H1009
4	133424	Ballari	Female	1979-06-11	+91 0106026841	Suicide	Bihar Sharif	H1017
5	172579	Devnath	Female	1982-02-22	+91 1868774631	Food allergy	Bidhannagar	H1019
6	171320	Atasi	Male	1980-02-21	+91 9747336855	Whiplash	Amravati	H1013
7	107794	Manish	Male	2010-01-26	+91 4354294043	Sunbathing	Panvel	H1004
8	130339	Aakar	Female	1990-04-24	+91 2777633911	Drug consumption	Bihar Sharif	H1000
9	110377	Gurudas	Male	1981-07-16	+91 1232859381	Dengue	Kamarhati	H1001
10	149367	NA	Male	1955-06-03	+91 1780763280	Head banging	Bangalore	H1013
11	156168	NA	Male	2010-02-17	+91 5586075345	Fanconi anaemia	Rajkot	H1004
	114241	NA	Female	2001-04-30	+91 4146391938	Breast cancer	Ghaziabad	H1015
13	146382	Dhamadaas	Male	1964-10-25	+91 6345482027	Anthrax	Bhalswa Jahangir Pu	H1019
14	132748	Brahmvir	Male	1944-04-04	+91 7316972612	Cystic fibrosis	Ambala	H1018
15	167340	NA	Female	1953-04-04	+91 2960004518	Galactosemia	Surendranagar Dudh	H1003
16	135184	Bhagvan	Female	2011-02-26	+91 0297693485	Dengue	Bhimavaram	H1018
17	179662	Amritkala	Female	1992-04-29	+91 0537157280	Smallpox	Meerut	H1018
18	184479	Bandhu	Male	1981-05-04	+91 0695289163	Pollen allergy	Chinsurah	H1010
19	156988	Bhagavaana	Female	1950-07-31	+91 6071745855	Breast cancer	Shahjahanpur	H1012
20	132870	NA	Female	1959-01-06	+91 8906694405	Glaucoma	Jabalpur	H1017
21	148137	Umang	Female	2017-02-26	+91 9485838770	Pet allergy	Haridwar	H1002
22	113280	Darsana	Male	1951-09-11	+91 7676311811	Rett Syndrome	Dibrugarh	H1019
23	134184	Prakash	Female	1998-06-26	+91 9268324471	Flu	Kottayam	H1001
24	122592	Vaijayanti	Male	1969-04-06	+91 9358851649	Cholera	Mira-Bhayandar	H1009



subscriber.csv

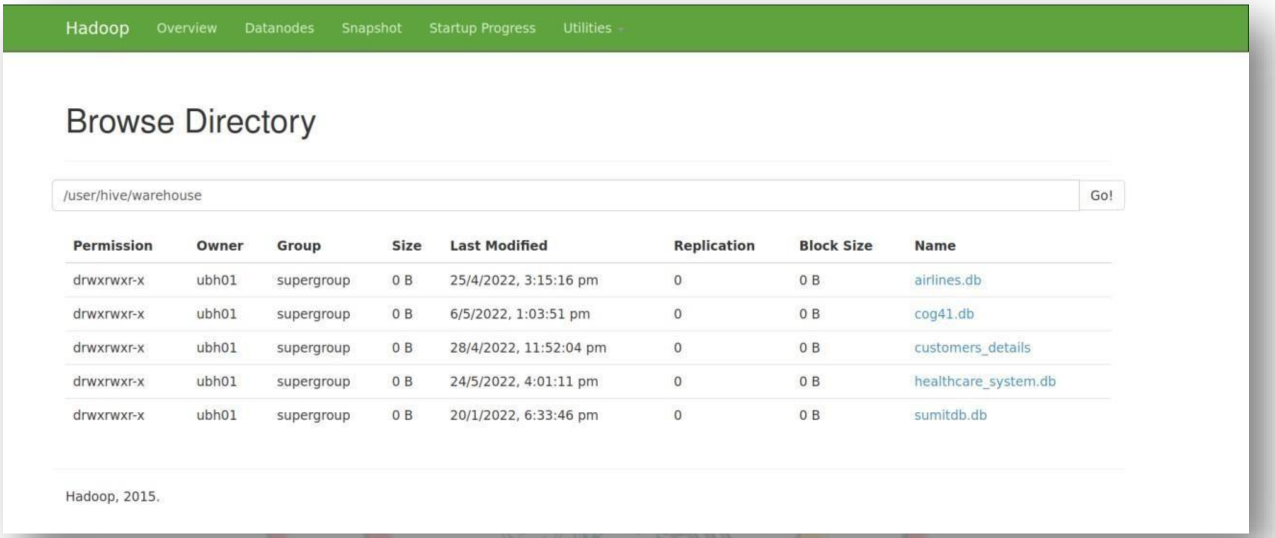
Open with

A	B	C	D	E	F	G	H	I	J	K	L	M
0	SUBID10000	Harbir	Vishwakarma	Baria Marg	1924-06-30	Female	+91 0112009318	India	Rourkela	767058	S107	Y
1	SUBID10001	Brahmdev	Sonkar	Lala Marg	1948-12-20	Female	+91 1727749552	India	Tiruvottiyur	34639	S105	Y
2	SUBID10002	Ujjawal	Devi	Mammen Zila	1980-04-16	Male	+91 8547451606	India	Berhampur	914455	S106	N
3	SUBID10003	Ballari	Mishra	Sahni Zila	1969-09-25	Female	+91 0106026841	India	Bihar Sharif	91481	S104	N
4	SUBID10004	Devnath	Srivastav	Magar Zila	1946-05-01	Female	+91 1868774631	India	Bidhannagar	531742	S110	N
5	SUBID10005	Atasi	Seth	Khatri Nagar	1967-10-02	Male	+91 9747336855	India	Amravati	229062	S104	Y
6	SUBID1006	Manish	Maurya	Swaminathan Chowk	1967-06-06	Male	+91 4354294043	India	Panvel	438733	S109	NA
7	SUBID10007	Aakar	Yadav	Swamy	1925-03-05	Female	+91 2777633911	India	Bihar Sharif	535907	S104	N
8	SUBID10008	Gurudas	Gupta	Sarin Nagar	1945-05-06	Male	+91 1232859381	India	Kamarhati	933226	S103	Y
9	SUBID10009	NA	Gupta	Thakur Circle	1925-06-12	Male	+91 1780763280	India	Bangalore	957469	S105	Y
10	SUBID1010	NA	Divedi	Dhillon	1976-02-03	Male	+91 5586075345	India	Rajkot	911319	S102	Y
11	SUBID10011	NA	Vishwakarma	Rajagopalan	1955-01-22	Female	+91 4146391938	India	Ghaziabad	337042	S106	N
12	SUBID10012	Dharmadaas	Tiwari	Rama	1964-04-29	Male	+91 6345482027	India	Bhalswa Jahangir Pu	430793	S103	N
13	SUBID10013	Brahmvir	Rai	Shah Path	1991-11-11	Male	+91 7316972612	India	Ambala	249898	S106	N
14	SUBID10014	NA	Srivastav	Chandra Path	1981-01-25	Female	+91 2960004518	India	Surendranagar Dudh	111966	S102	N
15	SUBID10015	Bhagvan	Srivastav	Edwin	1966-07-24	Female	+91 0297693485	India	Bhimavaram	436513	S105	Y
16	SUBID10016	Amritkala	Srivastav	Guha Path	1933-11-20	Female	+91 0537157280	India	Meerut	863467	S106	Y
17	SUBID10017	Bandhu	Seth	Varughese	1996-10-15	Male	+91 0695289163	India	Chinsurah	136713	S108	N
18	SUBID10018	Bhagavaana	Kumar	Kulkarni Zila	1935-09-16	Female	+91 6071745855	India	Shahjahanpur	597276	S101	N
19	SUBID10019	NA	Maurya	Sharaf Nagar	1924-11-09	Female	+91 8906694405	India	Jabalpur	958538	S104	N
20	SUBID1020	Umang	Srivastav	Balay Chowk	1963-07-14	Female	+91 9485838770	India	Haridwar	181692	S109	Y
21	SUBID1021	Darsana	Yadav	Upadhyay Zila	1932-05-29	Male	+91 7676311811	India	Dibrugarh	187414	S109	Y
22	SUBID10022	Prakash	Rao	Sachar	1923-09-15	Female	+91 9268324471	India	Kottayam	180680	S104	N
23	SUBID10023	Vaijayanti	Pratap	Khalsa Nagar	1920-11-13	Male	+91 9358851649	India	Mira-Bhayandar	419190	S102	Y

6.2 Hive and Sqoop

We have used Sqoop to import the data form RDBMS to Hive and there we can perform our necessary tasks to get the outputs

Here is the HEALTHCARE_SYSTEM Database created in Hive.



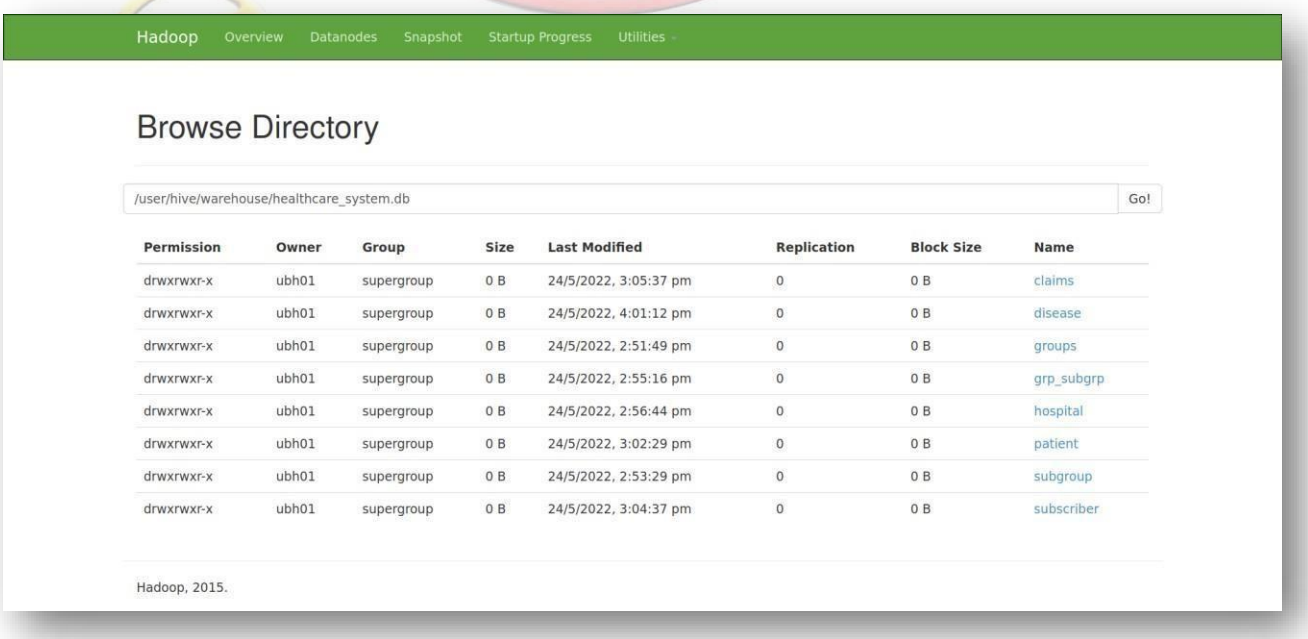
Browse Directory

/user/hive/warehouse

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxr-x	ubh01	supergroup	0 B	25/4/2022, 3:15:16 pm	0	0 B	airlines.db
drwxrwxr-x	ubh01	supergroup	0 B	6/5/2022, 1:03:51 pm	0	0 B	cog41.db
drwxrwxr-x	ubh01	supergroup	0 B	28/4/2022, 11:52:04 pm	0	0 B	customers_details
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 4:01:11 pm	0	0 B	healthcare_system.db
drwxrwxr-x	ubh01	supergroup	0 B	20/1/2022, 6:33:46 pm	0	0 B	sumitdb.db

Hadoop, 2015.

The tables created in the databases as mentioned in the schema



Browse Directory

/user/hive/warehouse/healthcare_system.db

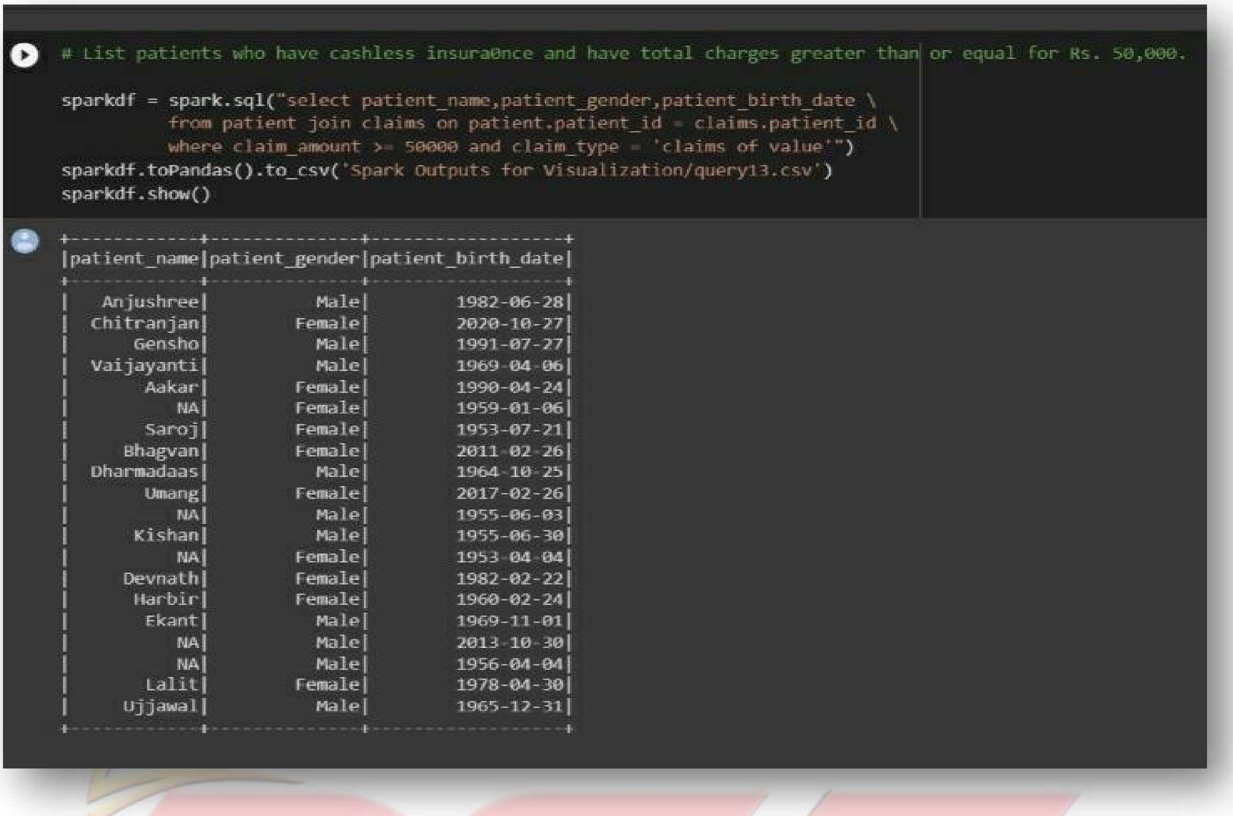
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 3:05:37 pm	0	0 B	claims
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 4:01:12 pm	0	0 B	disease
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 2:51:49 pm	0	0 B	groups
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 2:55:16 pm	0	0 B	grp_subgrp
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 2:56:44 pm	0	0 B	hospital
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 3:02:29 pm	0	0 B	patient
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 2:53:29 pm	0	0 B	subgroup
drwxrwxr-x	ubh01	supergroup	0 B	24/5/2022, 3:04:37 pm	0	0 B	subscriber

Hadoop, 2015.

6.3 Apache Spark

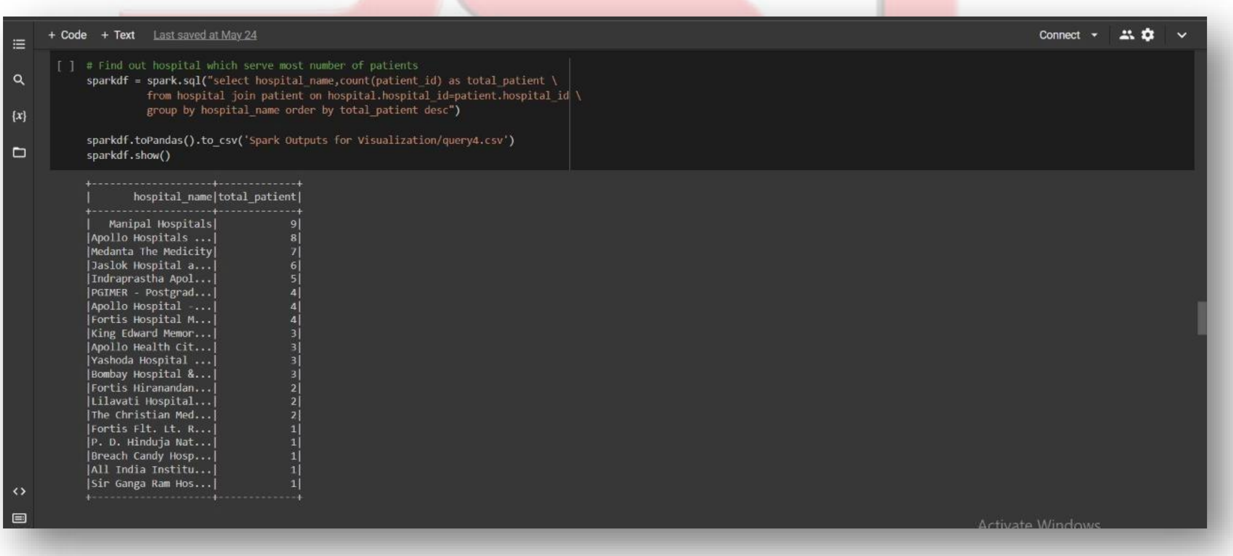
After uploading the data in to HDFS we connected spark. Here we analyze the data with help of python. Here we get our desired result in tabular form and that result is used to visualize our use cases.

Some snippet of the following code and result-



```
# List patients who have cashless insura0nce and have total charges greater than or equal for Rs. 50,000.

sparkdf = spark.sql("select patient_name,patient_gender,patient_birth_date \
          from patient join claims on patient.patient_id = claims.patient_id \
          where claim_amount >= 50000 and claim_type = 'claims of value'")
sparkdf.toPandas().to_csv('Spark Outputs for Visualization/query13.csv')
sparkdf.show()
```

```
+------------+--------------+------------------+
|patient_name|patient_gender|patient_birth_date|
+------------+--------------+------------------+
|   Anjushree|          Male|        1982-06-28|
|  Chitranjan|        Female|        2020-10-27|
|      Gensho|          Male|        1991-07-27|
|  Vaijayanti|          Male|        1969-04-06|
|       Aakar|        Female|        1990-04-24|
|          NA|        Female|        1959-01-06|
|       Saroj|        Female|        1953-07-21|
|     Bhagvan|        Female|        2011-02-26|
|  Dharmadaas|          Male|        1964-10-25|
|       Umang|        Female|        2017-02-26|
|          NA|          Male|        1955-06-03|
|      Kishan|          Male|        1955-06-30|
|          NA|        Female|        1953-04-04|
|     Devnath|        Female|        1982-02-22|
|      Harbir|        Female|        1960-02-24|
|       Ekant|          Male|        1969-11-01|
|          NA|          Male|        2013-10-30|
|          NA|          Male|        1956-04-04|
|       Lalit|        Female|        1978-04-30|
|     Ujjawal|          Male|        1965-12-31|
+------------+--------------+------------------+
```



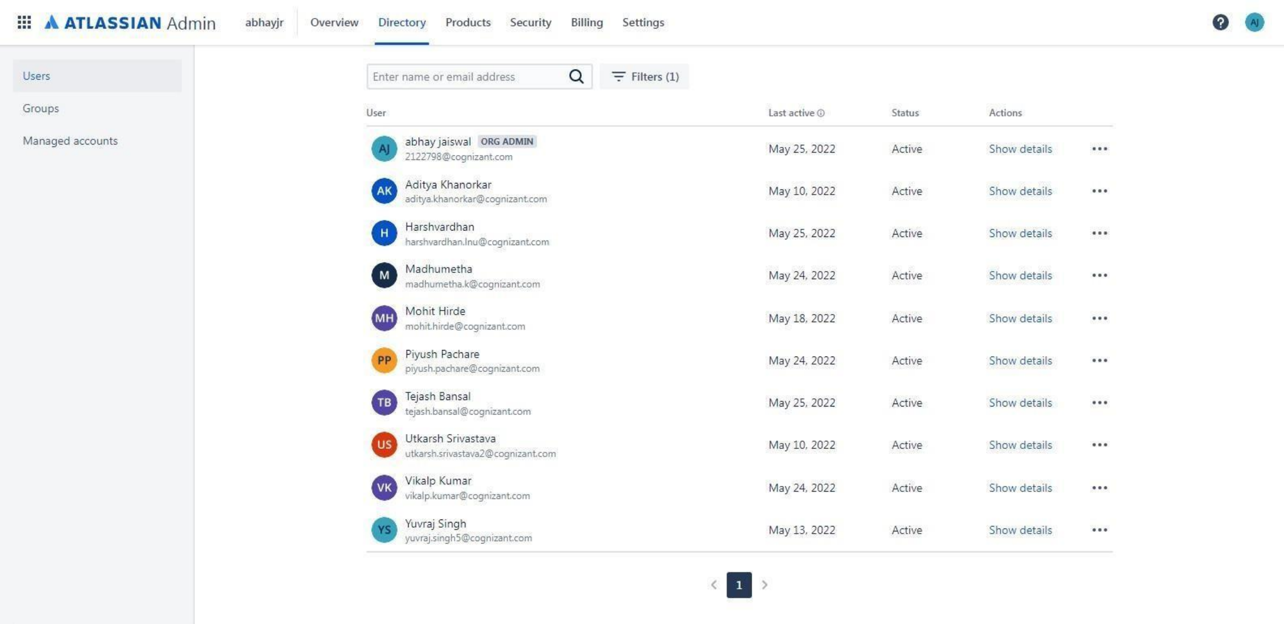
```
[ ] # Find out hospital which serve most number of patients
    sparkdf = spark.sql("select hospital_name,count(patient_id) as total_patient \
              from hospital join patient on hospital.hospital_id=patient.hospital_id \
              group by hospital_name order by total_patient desc")

    sparkdf.toPandas().to_csv('Spark Outputs for Visualization/query4.csv')
    sparkdf.show()
```

```
+--------------------+-------------+
|       hospital_name|total_patient|
+--------------------+-------------+
|   Manipal Hospitals|            9|
|Apollo Hospitals ...|            8|
|Medanta The Medicity|            7|
|Jaslok Hospital a...|            6|
|Indraprastha Apol...|            5|
|PGIMER - Postgrad...|            4|
|Apollo Hospital -...|            4|
|Fortis Hospital M...|            4|
|King Edward Memor...|            3|
|Apollo Health Cit...|            3|
|Yashoda Hospital ...|            3|
|Bombay Hospital &...|            3|
|Fortis Hiranandan...|            2|
|Lilavati Hospital...|            2|
|The Christian Med...|            2|
|Fortis Flt. Lt. R...|            1|
|P. D. Hinduja Nat...|            1|
|Breach Candy Hosp...|            1|
|All India Institu...|            1|
|Sir Ganga Ram Hos...|            1|
+--------------------+-------------+
```

7. Project Management Tool

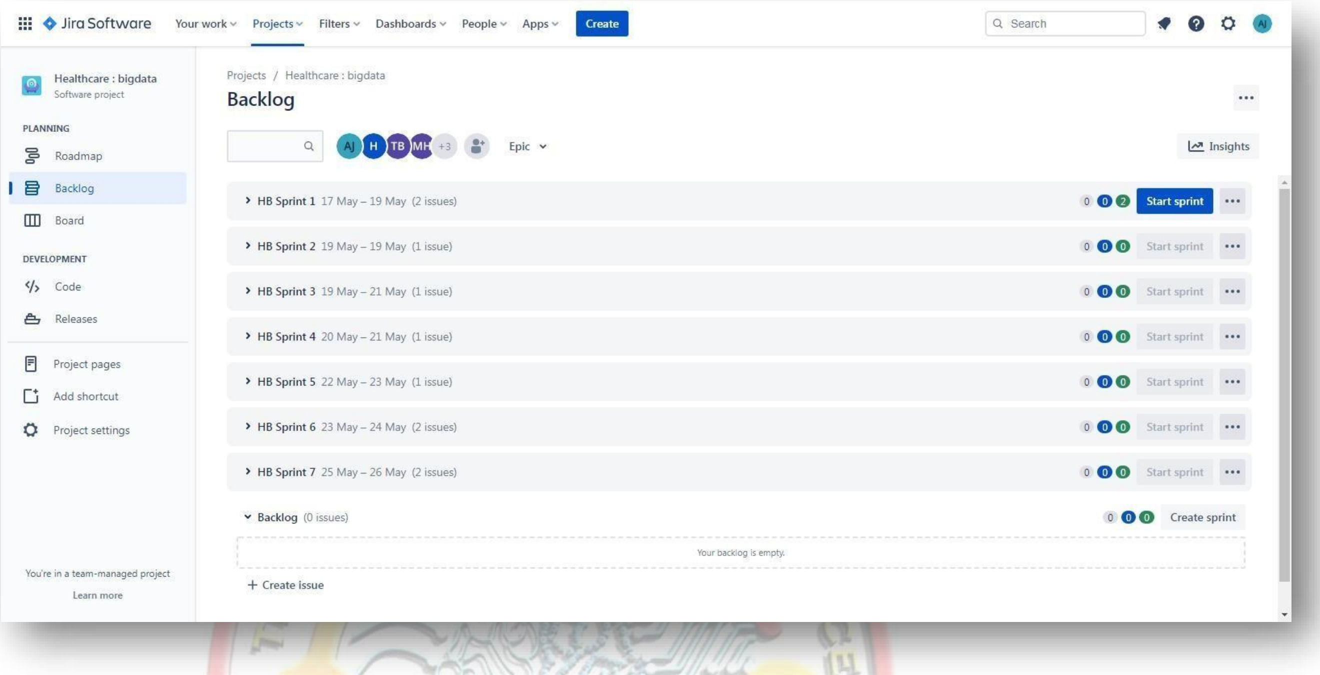
Jira Software is part of a family of products designed to help teams of all types manage work. Originally, Jira was designed as a bug and issue tracker. But today, Jira has evolved into a powerful work management tool for all kinds of use cases, from requirements and test case management to agile software development.

In this project we use Jira as a Project Management tool. With the help of the Jira, we assign a task and customizes the issues and subtasks in a whole team easily, also manages the workflow and track the progress. It also helps us to change the permission for a particular task within the team.



User	Last active	Status	Actions
abhay jaiswal ORG ADMIN 2122798@cognizant.com	May 25, 2022	Active	Show details •••
Aditya Khanorkar aditya.khanorkar@cognizant.com	May 10, 2022	Active	Show details •••
Harshvardhan harshvardhan.lnu@cognizant.com	May 25, 2022	Active	Show details •••
Madhumetha madhumetha.k@cognizant.com	May 24, 2022	Active	Show details •••
Mohit Hirde mohit.hirde@cognizant.com	May 18, 2022	Active	Show details •••
Piyush Pachare piyush.pachare@cognizant.com	May 24, 2022	Active	Show details •••
Tejash Bansal tejash.bansal@cognizant.com	May 25, 2022	Active	Show details •••
Utkarsh Srivastava utkarsh.srivastava2@cognizant.com	May 10, 2022	Active	Show details •••
Vikalp Kumar vikalp.kumar@cognizant.com	May 24, 2022	Active	Show details •••
Yuvraj Singh yuvraj.singh5@cognizant.com	May 13, 2022	Active	Show details •••

ASSURANCE OF EXCEL PST IT SOLUTIONS



Projects / Healthcare : bigdata

Backlog

HB Sprint 1 17 May – 19 May (2 issues) 0 0 2 Start sprint

HB Sprint 2 19 May – 19 May (1 issue) 0 0 0 Start sprint

HB Sprint 3 19 May – 21 May (1 issue) 0 0 0 Start sprint

HB Sprint 4 20 May – 21 May (1 issue) 0 0 0 Start sprint

HB Sprint 5 22 May – 23 May (1 issue) 0 0 0 Start sprint

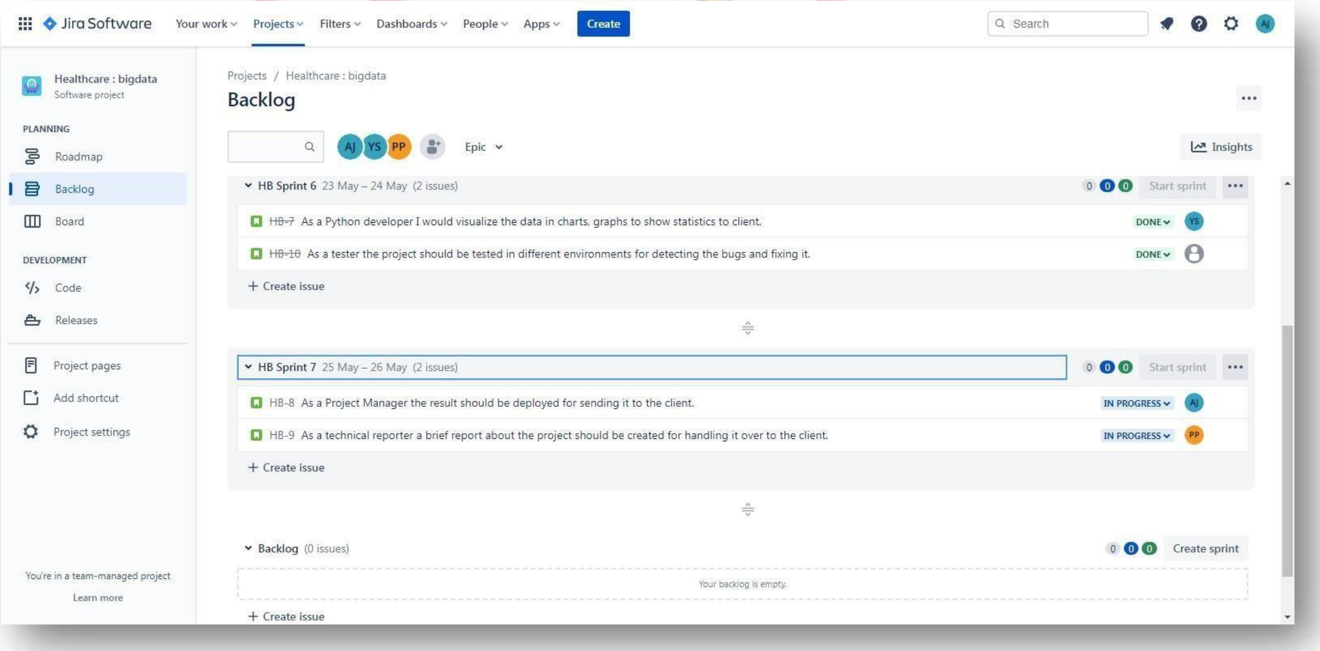
HB Sprint 6 23 May – 24 May (2 issues) 0 0 0 Start sprint

HB Sprint 7 25 May – 26 May (2 issues) 0 0 0 Start sprint

Backlog (0 issues) 0 0 0 Create sprint

Your backlog is empty.

+ Create issue



Projects / Healthcare : bigdata

Backlog

HB Sprint 6 23 May – 24 May (2 issues) 0 0 0 Start sprint

~~HB-7~~ As a Python developer I would visualize the data in charts, graphs to show statistics to client. DONE

~~HB-10~~ As a tester the project should be tested in different environments for detecting the bugs and fixing it. DONE

+ Create issue

HB Sprint 7 25 May – 26 May (2 issues) 0 0 0 Start sprint

HB-8 As a Project Manager the result should be deployed for sending it to the client. IN PROGRESS

HB-9 As a technical reporter a brief report about the project should be created for handling it over to the client. IN PROGRESS

+ Create issue

Backlog (0 issues) 0 0 0 Create sprint

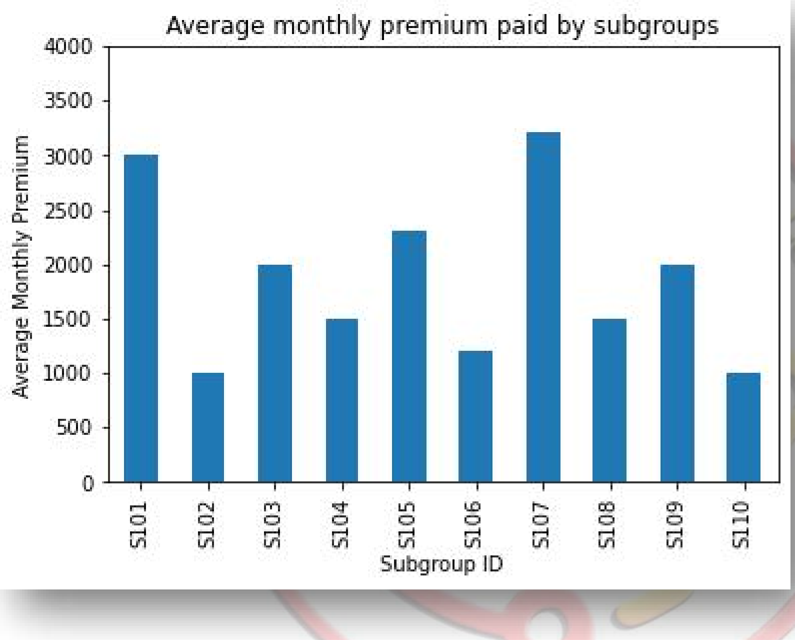
Your backlog is empty.

+ Create issue

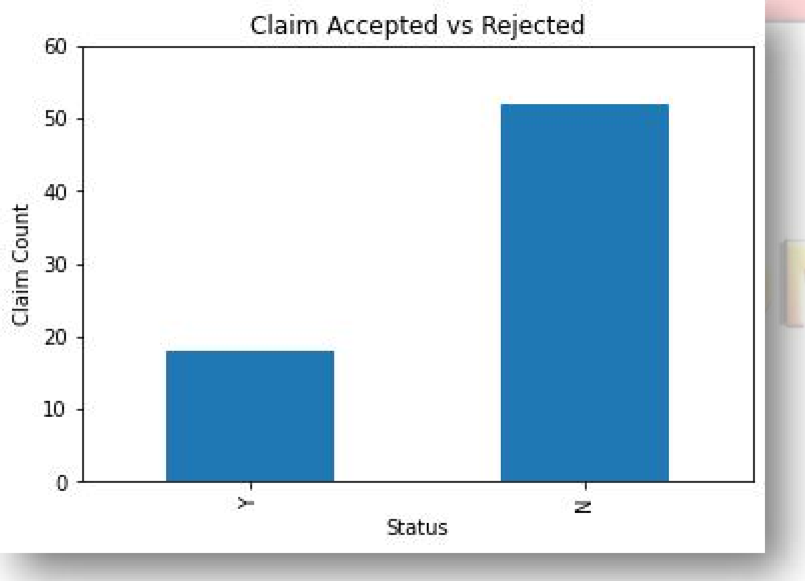
8. Output Screens

We used Matplotlib and seaborn to visualize our use cases which will be better to take business decision.

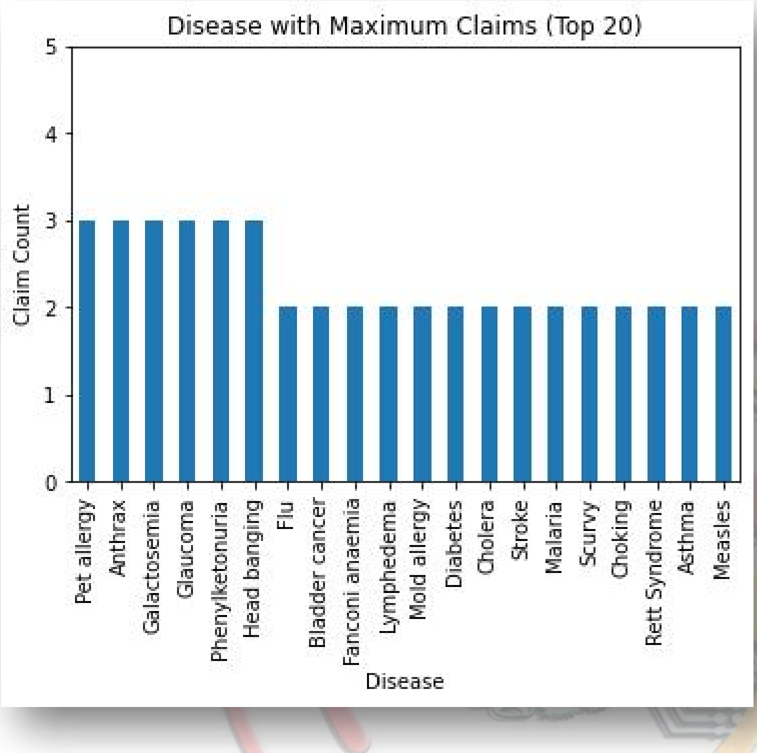
Use Case-1: Average Monthly premium for each subgroup



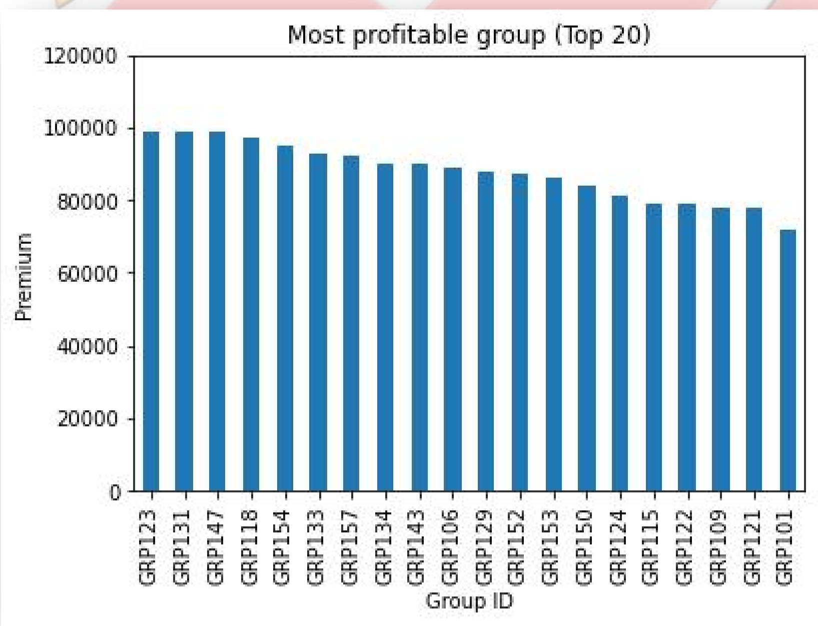
Use Case-2: Number of people whose claim either got accepted or rejected.



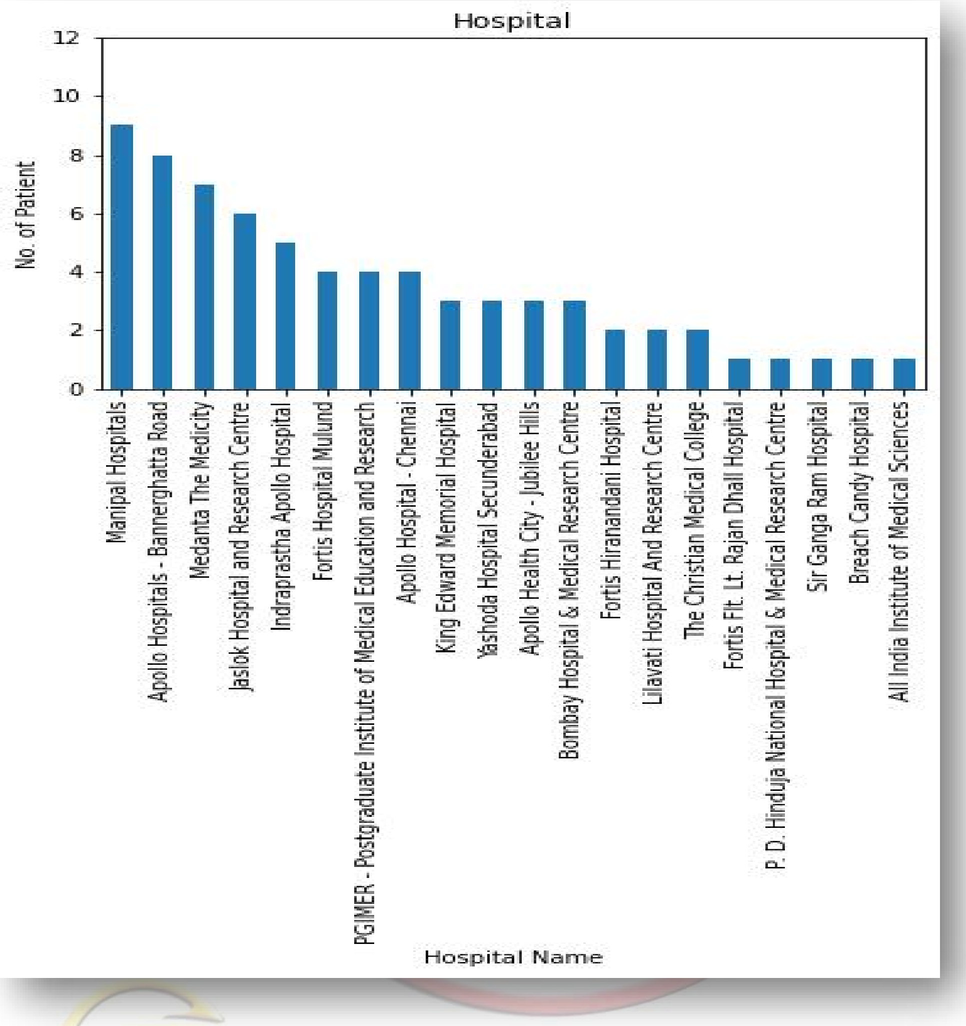
Use case-3: Which disease have maximum number of claims



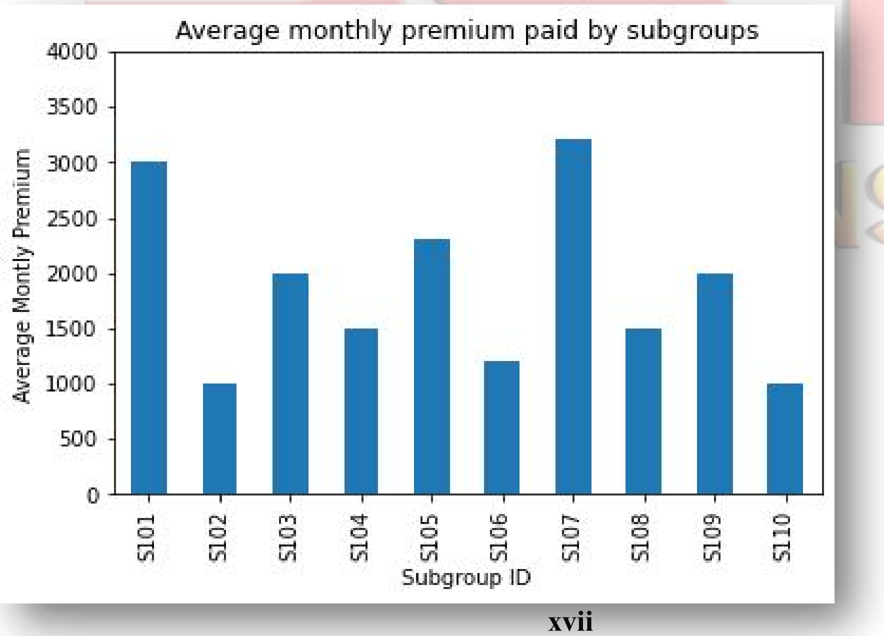
Use Case-4: Which company/group is most profitable



Use case-5: No. of patient in each hospital



Use case-6: Average Monthly premium paid by each subgroup.



9. Conclusion

We have collected data from various 3rd party sources and processed them and with the help of Big Data tools we computed the data to visualize some of necessary use case. Based on theabove analysis the health care insurance company will create a new business strategy to acquire more customers, engagement and send offers. As well as fetching the company and customer details and provide easy access to information regarding customers.

10. Further Enhancements/Recommendations

This project has a very vast scope in future in this field. We developed this project on the requirement of our client but it can be generalized in future. If we get required resources, wecan get more accurate results. There are various use cases that can be achieved by this project. Some of future scopes are bellow-

- Real time data can also be used for real time processing.
- We can automate the whole procedure where data coming from sources and gettingexecuted at a same time.
- Not in the Healthcare industry we can generalized the whole procedure to othersectors like cars, online education system etc.

What is *Attunity?*

Attunity is now part of Qlik.

Qlik Data Integration efficiently delivers large volumes of real-time, analytics-ready data into streaming and cloud platforms, data warehouses, and data lakes. And with an *agentless and log-based approach to change data capture, your data is always currentwithout impacting source systems.*

For more info:- https://www.qlik.com/us/streaming-data/data-streaming-cdc

What is SFTP?

SFTP is the abbreviation of Secure File Transfer Protocol. It is a file protocol used totransfer large files over the Web. It is built on file transfer protocol (FTP) and includes Secure Shell (SSH) security components.

Secure Shell is an encryption component for Internet security. SSH and SFTP were designed by the Internet Engineering Task Force (IETF) to improve web security. SFTP uses SSH and encrypted FTP commands to transfer files securely to avoid password sniffing and exposing sensitive information in plain text. SFTP can also prevent man-in-the-middle attacks since the client needs to be authenticated by the server.

How Does SFTP Work?

How does SFTP work? The SFTP establishes a secure connection through an SSH datastream and provides organizations with a higher level of file transfer protection. This is because SFTP uses encryption algorithms to safely move data to your server and keep files unreadable during the process, and authentication prevents unauthorized file access during operations.

Although the SFTP does not require two-factor authentication, you can choose to require both a user ID and password as well as an SSH key for a more secure connection. Creating SSH keys helps prevent imposters from connecting to the SFTPserver. The SSH key pair must be generated in advance

For More Info :- https://www.minitool.com/lib/what-is-sftp.html

What is Connect Direct Gateways?

Connect Direct—originally named Network Data Mover (NDM)— is a computer softwareproduct that **transfers files between mainframe computers and/or midrange computers**. It was developed for mainframes, with other platforms being added as the product grew.

It is point-to-point (peer-to-peer) file-based integration middleware meant for 24x7 unattended operation, which provides assured delivery, high-volume, and secure data exchange within and between enterprises. It is optimized for high performance and throughput and moves files containing any type of data (text, EDI, binary, digital content, image) across multiple platforms, disparate file systems, and disparate media.It is used by many industries throughout the world to move large volumes of data and for connecting to remote offices.

What is Control-M Tool?

Control-M **simplifies application and data workflow orchestration on premises or as aservice**. It makes it easy to build, define, schedule, manage, and monitor productionworkflows, ensuring visibility, reliability, and improving SLAs.

Alternatives Schedulers: - Apache Oozie, Snowflake, IBM Workload Automation

Cluster Configuration:-

Size of Data Node = 20TB

Data Node(Node

manager) = 17Name

Node (active) = 1

Name Node

(Passive) = 1

Secondary Name

Node = 1Resource

Manager = 1

Total Node in Cluster => 21 (Number of nodes are odd always)

Some Project based Interview Questions :-

Q1. How do you validate data after data ingestion/migration?

Number of record
Column property map or not
Datatype

Q2. What is your Daily data intake?

-> 15 to 20GB

Q3. How many number of tables do you worked with?

Total Tables -> 200

70 - daily

30 - weekly

20 - monthly

Q4. What Compression Tech you are using in your project?

Snappy (default)

Q5. What are some common challenges you may encounter **during data migration?**

Some of the more common challenges include:

Large Data Set Size: Data from many tables may need to be migrated, which couldincrease the complexity of a project.

Complex Data Relationships: Different types of relationships may exist betweendifferent tables and columns, requiring more effort during the migration.

Data validation: Ensuring that all the data is valid and correct before migrating it can bedifficult, especially if you migrate large amounts of data.

Data inconsistency: It's vital to ensure that the source and destination systems have similar naming conventions. Otherwise, there can be problems during or after migrationwith differing data names.

Target format differences: If your target environment is running on another platform, it may not support all of the same data formats as the source system. It can cause problems during or after a migration.

Different Data Quality: Data quality may differ between source and target data, makingit harder to match records.

What is ur roles and responsibilities?

The data migration team has several roles and responsibilities. our primary responsibility is to ensure the data is extracted from sources and moved to the databasein an efficient and trustworthy manner. The duties the team performs include extracting data from the original sources, assuring the quality and cleaning the data, applying labels and measures to the data, and delivering the data in a format that can be used by query tools, report writers, and dashboards.

What is the best way to process a fixed-length flat file?

Fixed-length flat files are simple to process if you first define the layout of the files. The information included in this definition is usually the beginning of the file, its length, and the data type. This information is initially entered manually and will remain the same until the data migration tool encounters a file with a different format. This will cause the migration to fail and alert you that some manual intervention is required. It is recommended that you perform periodic validation processes to ensure that the data format has not changed. . Data validation is essential in this process in case the formatof the data has changed. The most likely change will occur if the data includes a date field

What is the difference between an initial load and a full load process in the context ofdata migration?

In the context of data migration, the initial load is the process for populating all data warehousing tables for the first time. Full load also refers to the first time a data warehouse is populated. However, using the full load process, all the records are loadedin one batch after all the contents of the table have been erased."

Q. what’s your data source?

In which format your data was?(File formats,nature of data)

Q. what's your cluster size.

Ans. We have a 21 node Cluster.

In future it might grow based on the needs.

Q. Kindly explain your project architecture(end to end

explanation)(It should be more than 10 mints)

Q. How much new data you get on

daily basis.Ans. 15-20 GB Per day

Incremental Data.

Also we expect the data to grow over time.

Q. what is your role in your big data project.

Ans. I am involved in ingesting the data.

I majorly work on Hive, Sqoop & their Performance Tuning aspects.

Q. which big data distribution are you using.

Ans. Before few years ago we used Cloudera distribution package and now we areplanning to move on Amazon web services.

Q. what's the configuration of each node in your cluster.

Ans. 64 GB RAM & 16 CPU Cores each.

Q. Did you ever face any performance challenges with your job? how did you optimizethat.

Ans. Yes, Many a times.

sometimes we see job taking lot of time, due to lot of shuffling and we try to minimize it.

Q. What optimisation techniques have you used in your project?

Ans:- Performance a)Handle data skewness b)Use filter first then sort while doing joinsspecially “Hive”:-a) Dynamic Partitions ,b)Bucketing ,c)External table

Q. If you are facing with an issue while execution of bash file with sql script in hive - italways stuck at the same place map=100%, reduce=67%

The phases of a Reducer are:

- Shuffle

- Sort
- Reduce

Getting stuck @ 67% indicates that the Shuffle and Sort have completed but none of your partitions are able to succeed in the Reduce phase. The Reduce phase is your actual Reducer code. This indicates your code is unable to complete. You should examine your code and also look at the hive logs to see why your code is unable to be run.

And also try with these options with mappers and reducers number with differentvariations and other tunning characteristics like :-

```
SET hive.exec.parallel=true;
SET
hive.default.filefor
mat=RCFILE;SET
hive.stats.autogath
er=false;
SET hive.exec.compress.output=true;
SET
mapred.output.compression.codec=org.apache.hadoop.io.compre
ss.SnappyCodec;SET mapred.output.compression.type=BLOCK;
SET
hive.input.format=org.apache.hadoop.hive.ql.io.CombineHiveInputFormat;.
```

Q. How you implemented integration with different different module?

Q. which version of modules you used in your project?

Q. What is the most challenging problem you have solved in your big data project?

Q. When would you prefer to use Hive, and when would you prefer Spark SQL?

Sqooping issues:

1) Data type conversion issue: will have to be very careful when we import the data from RDBMS to hadoop system , you will notice default conversion happening and are not suitable to the business need. we can use map-column-java/map-column-hive function tohandle this issue

2)Sqoop does not support few hadoop file types like ORC,RC,Parquet

3)Delimiters : make sure the delimiter you are using is not part of the data you areimporting/loading

4)import-all command will not work if your tables are not having primary key for sqoop tosplit the data

5)if you need multiple mappers(parallel import) you will have to provide split-by column

6)Table and column names can not have special characters . When importing data from legacy system you might face this issue, so we will have to write scripts to take care of this.

OutOfMemoryError in Sqoop and Hive :-

The OutOfMemoryError exception usually happens during **INSERT OVERWRITEcommands when there's not enough heap space** on hive-server2, the Hive metastore, or the client side. To resolve this issue, increase the maximum memory allocation for the JVM or increase HADOOP_HEAPSIZE.

Sqoop mappings sometimes fail with an out-of-memory error on the Spark engine. This issue occurs if the **Java heap**

size is not sufficient. Solution To resolve this issue, increase the Java heap size value for the Data IntegrationService that runs the Sqoop mapping, and then run the mapping again.

Refer Interview questions :-

Top 50 Apache Sqoop Interview Questions & Answers - DataFlair (data-flair.training)

Tricky Hive Interview Questions and Answers for Experience - DataFlair (data-flair.training)

Top 30 Tricky Hive Interview Questions and Answers - DataFlair (data-flair.training)

Top 30 Hive Interview Questions & Answers 2022 - Intellipaa