Nitte(DU) established under Section 3 of UGC Act 1956 | Accredited with ‘A+’ Grade by NAAC 🕿:

08258 - 281039 – 281263, Fax: 08258 – 281265

**HEALTHCARE SYSTEM**

***A Mini Project Report Submitted by***

**MEGHANA BHAT PERVAJE KAVANA PAI (4NM20AI025) (4NM20AI020)**

**UNDER THE GUIDANCE OF** 

**Ms.Rakshitha**

**Assistant Professor Gd-1**

**Department of Artificial Intelligence and Machine Learning Engineering**

***In partial fulfillment of the requirements for the***

***BIG DATA ANALYTICS(20AI505)***

****December 2020**

****(ISO 9001:2015 Certified), Accredited with ‘A’ Grade by NAAC

🕿: 08258 - 281039 – 281263, Fax: 08258 – 281265

**Department of Artificial Intelligence and Machine Learning Engineering**

B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

CERTIFICATE

*Certified that the mini project work entitled*

***“Healthcare System”***

*is a bonafide work carried out by*

***Meghana Bhat Pervaje Kavana Pai (4NM20AI025) (4NM20AI020)***

*in partial fulfilment of the requirements for the award of*

***Bachelor of Engineering Degree*** *in* ***Artificial Intelligence and Machine Learning Engineering*** *prescribed by* ***Visvesvaraya Technological University, Belgaum***

*during the year 2022-2023.*

*It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library.*

*The mini project report has been approved as it satisfies the academic requirements in respect of the mini project work prescribed for the Bachelor of Engineering Degree.*

**Signature of Guide Signature of HOD**

**Evaluation**

**Name of the Examiners Signature with Date**

*1.*

*2.*

**ACKNOWLEDGEMENT**

We believe that our mini project will be complete only after we thank the people who have contributed to make this mini project successful.

First and foremost, our sincere thanks to our beloved principal, **Dr. Niranjan N. Chiplunkar** for giving us an opportunity to carry out our mini project work at our college and providing us with all the needed facilities.

I acknowledge the support and valuable inputs given by**, Dr.Sharada U Shenoy** the Head of the Department, Artificial Intelligence and Machine Learning Engineering, NMAMIT, Nitte

We express our deep sense of gratitude and indebtedness to our guide **Ms.Rakshitha,** Assistant Professor Artificial Intelligence and Machine Learning Engineering, for his inspiring guidance, constant encouragement, support and suggestions for improvement during the course for our mini project.

We also thank all those who have supported us throughout the entire duration of our mini project.

Finally, we thank the staff members of the Department of Artificial Intelligence and Machine Learning Engineering and all our friends for their honest opinions and suggestions throughout the course of our mini mini project.

**Meghana Bhat Pervaje**

**Kavana Pai**

iv

**ABSTRACT**

Big data’ is massive amounts of information that can work wonders. It has become a topic of special interest for the past two decades because of a great potential that is hidden in it. Various public and private sector industries generate, store, and analyze big data with an aim to improve the services they provide. In the healthcare industry, various sources for big data include hospital records, medical records of patients, results of medical examinations, and devices that are a part of internet of things. Biomedical research also generates a signifcant portion of big data relevant to public healthcare. This data requires proper management and analysis to derive meaningful information. Otherwise, seeking solution by analyzing big data quickly becomes comparable to finding a needle in the haystack. There are various challenges associated with each step of handling big data which can only be surpassed by using high-end computing solutions for big data analysis. That is why, to provide relevant solutions for improving public health, healthcare providers are required to be fully equipped with appropriate infrastructure to systematically generate and analyze big data. An efcient management, analysis, and interpretation of big data can change the game by opening new avenues for modern healthcare. That is exactly why various industries, including the healthcare industry, are taking vigorous steps to convert this potential into better services and financial advantages. With a strong integration of biomedical and healthcare data, modern healthcare organizations can possibly revolutionize the medical therapies and personalized medicine.

v

**TABLE OF CONTENTS**

**Title Page....................................................................................................................i Certificate ..................................................................................................................ii Acknowledgement...................................................................................................iii Abstract....................................................................................................................iv Table of Contents .....................................................................................................v**

**Contents**

1 Business Challenge / Requirement………………………………………………………………... ii 2 Goal of the Project………………………………………………………………………………… ii 3 Data Flow Architecture / Process Flow…………………………………………………………... ii

3.1 Project Architecture………………………………………………………………….............. iii 4 Dataset Explanation ……………………………………………………………………................ iv 5 Problem Statement / Tasks………………………………………………………………………... vi

5.1 Problem Statement 1 …………………………………………………………………………...vi 5.2 Problem Statement 2…………………………………………………………………………. vii 6 Coding/Code Templates……………………………………………………………………………ix 6.1Data Processing………………………………………………………………………………… ix 6.1.1 Conversion of raw data to processed data………………………………………………. ix 6.1.2 Processed Dataset………………………………………………………………………….x

6.2 Hive and Sqoop………………………………………………………………………………. xi 6.3 Apache Spark…………………………………………………………………………………. xii 7 Project Management Tool………………………………………………………………………... xiii 8 Output Screens….………………………………………………………………………………... xv 9 Conclusion….……………………………………………………………………………………. xviii 10 Further Enhancements/Recommendations…………………...………………………………… xviii 11 References/Bibliography………………………………………………………………...............xviii

**i**

**1. 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.

**2. 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.

**3. Data Flow Architecture / Process Flow**

1. A Linux file server receives data files in form of json and csv. These files are coming from the third-party sources based on user interaction methodology.

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

3. After validated the files data, we are creating the data model for RDBMS so that we can store the files data into the RDBMS.

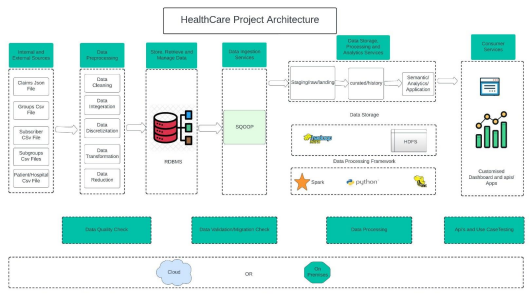
4. After storing the data into the RDBMS, we now transform according to our business requirements.

5. Finally, data landed to 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

**ii**

**3.1 Project Architecture**

**iii**

**4. 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

● 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

**iv**

**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

**CSV File 7 Fields (grpsubgrp.csv)**

● SUBGRP\_ID

● GRP\_ID

**v**

**5. 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.

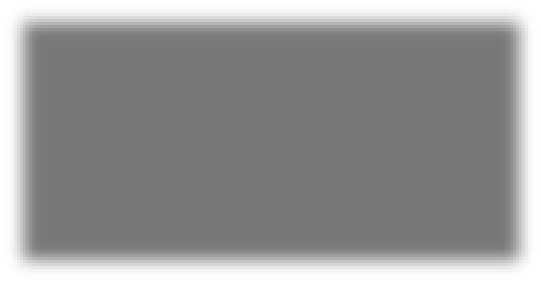
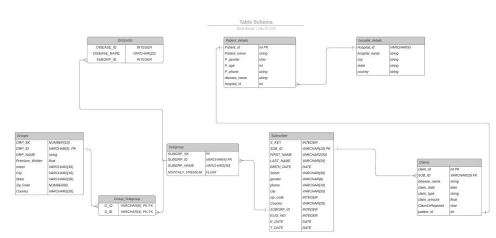
• Load only the members who are currently effective. (i.e.) SYSDATE BETWEEN EFFT\_DT AND TERM\_DT

• 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**

**vi**

**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

**vii**

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.

Additional Info

● Submitted code can run in any given Hadoop cluster of the same Hadoop version. Deliverables

Below are the expected deliverables-

● Code and link to code repository **GitHub Repo**

● Jupiter notebook/ VS code to run the code

**●** Any other script/wrapper not required to run the code in any environment

**viii**

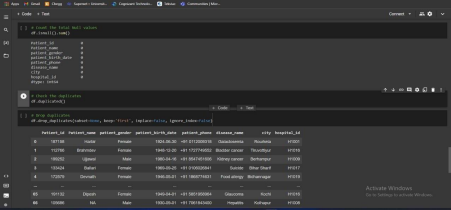
**6. Coding/Code Templates:**

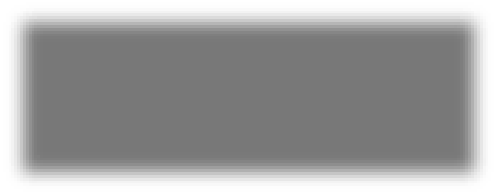
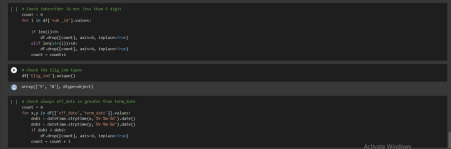
**6.1 Data Processing**

**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

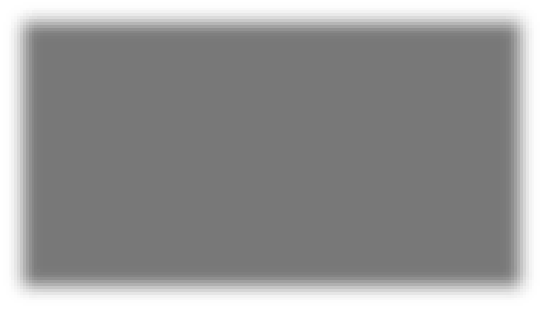
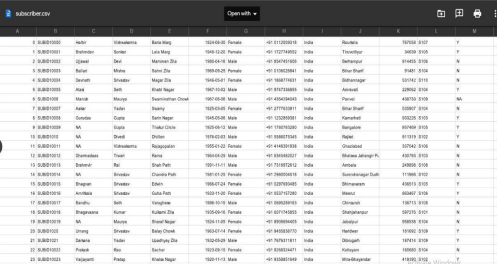
processed dataset. here are some samples of codes.

Here we are checking for SUB\_ID if is it length of 9 or not

**ix**

**6.1.2 Processed Dataset**

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

**x**

**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.



The tables created in the databases as mentioned in the schema

**xi**

**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-

**xii**

**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.

**xiii**

**xiv**

**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.**

**xv**

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

**Use Case-4: Which company/group is most profitablexvi**

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

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

**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 the above 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, we can 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 getting executed at a same time.

• Not in the Healthcare industry we can generalized the whole procedure to other sectors like cars, online education system etc.

**11. References/Bibliography**

Books and Reports

**[1] Beranger, Jérôme. 2016.** *Ethics in Big Data: the medical datasphere.* **London: Elsevier. [2]** Davis, Cord and Patterson, Doug. 2012. Ethics of Big Data. Farnham, O’Reilly.

Web Sites

**[1]** Big Data Ethics

[2] GeekforGeeks

**xviii**