

Big Data Analysis of Healthcare Data - Team9

Spark

Lakshmi Sireesha Pavalla S543026, Anudeep Somarouthu S545543, Niharika Sanamsetty S545249, Swapna Dasari S544719, rajashekhar Kota S545008.

Northwest Missouri State University, Maryville MO 64468, USA
S543026@nwmissouri.edu, S545543@nwmissouri.edu, S545249@nwmissouri.edu, S544719@nwmissouri.edu, S545008@nwmissouri.edu

1 Introduction Of Project

Here we would like to analyse the data of hospitals located in California in USA in various aspects. In the first place we would like to clean the data and also format the unstructured data to a structured data. Using the tools and technologies like databricks, PySpark, SQL, Hadoop MapReduce would like to analyse the goals.

1.1 Explanation Healthcare data set in terms of Big Data

1. Volume: By Using big data analytics we are processing California state hospital quarterly financial and utilization data files, data contains following records, quarterly data file labels quarters ended 2015 and after, it has records quarterly report information, general hospital information, utilization data, hospital discharges, patient days, outpatient visits, gross inpatient revenue, gross outpatient revenue, deductions from revenue, capitation premium revenue, net patient revenue, other revenue and expense data, purchased inpatient services, purchased outpatient services, other financial data items, covered California, quality assurance fee program, we process the data with all these records.
2. Velocity: we are expecting the data can be processed, stored and analyzed by relational database by increasing speed, we are going to observe the speed at which new data is generated and the speed at which data moves around.
3. Variety: The variety of structured and unstructured data increases the complexity of storing and analyzing data. we are generating data that is in unstructured form.
4. Veracity: The accuracy of data analysis depends on the veracity of the data, we make sure the data doesn't contain any duplicate data, we also make sure the data going to be accurate by removing unwanted data.
5. Value: It is one of the important aspects in bigdata, with this California state hospital data, one can easily access information about hospitals which is useful for both the faculty and patients

1.2 Goals Of The Project

1. In which locality highest no of health care units are observed?

2. In which locality lowest no of health care units are observed?
 3. Display the number of healthcare centers in every City?
 4. Segregate the count of hospital list depending upon the central type of each state How many are Non-Profit Corp? How many are Church?
 5. For which hospital highest no of available beds are recorded?
 6. For which hospital least no of available beds are recorded?
 7. Display the count of distinct HSA on overall health care data
 8. Categorize the hospital based on its type How many are Psychiatric Health Facilities? How many are Comparable?
 9. Highest capacity of licenced beds for hospitals
 10. lowest capacity of licenced beds for hospitals
- Based on this data we will have a clear picture of how strong or weak california region exists in healthcare.

1.3 Explanation of Tools and Technologies used in the project

Databricks,PySpark,SQL and Hadoop MapReduce.

1. Data bricks: Data bricks is used to process, store, clean, share, analyze, model data sets, by using data bricks we have extracted following results from our data set which are highest and lowest capacity of beds,highest and lowest count of beds,highest no of health care units.
2. PySpark: PySpark is an open source, distributed computing platform and collection of tools for real-time, massive data processing.with Pyspark we have shown some visualizations with our data.
3. SQL: By Using SQL statements we have obtain results from our dataset
4. Hadoop MapReduce: A software framework named Hadoop MapReduce makes it simple to create applications that efficiently handle huge amounts of data in parallel on large clusters of affordable hardware.By Using this we have extracted count of non profit corporation and psychiatric health facilities and how many are comparable.

1.4 Block diagram

We have chosen the data related to Health care which is unstructured data In order to transform the data into structured data we would be doing the data cleaning.Data cleaning includes based on the goals we performs we have removed the unnecessary columns and remove null values. Once the data is clean the xlsx file is converted to .CSV file to use the data to upload the data file to databricks so and so forth to perform various operations like order by and group by and other predefined functions like MAX and MIN based the goals we have defined to implement on the data.In the Databricks tool we would be using technologies like SQL,pyspark to extract the meaningful data. Part of the data of healthcare is used as input to the mapReduce function to get the results accordingly.We are with the results and goal here comes the Analysis part.Based on the results by using the SQL, PySpark and mapreduce technologies we would suggest and conclude and can make decision which is valuable and worthy.

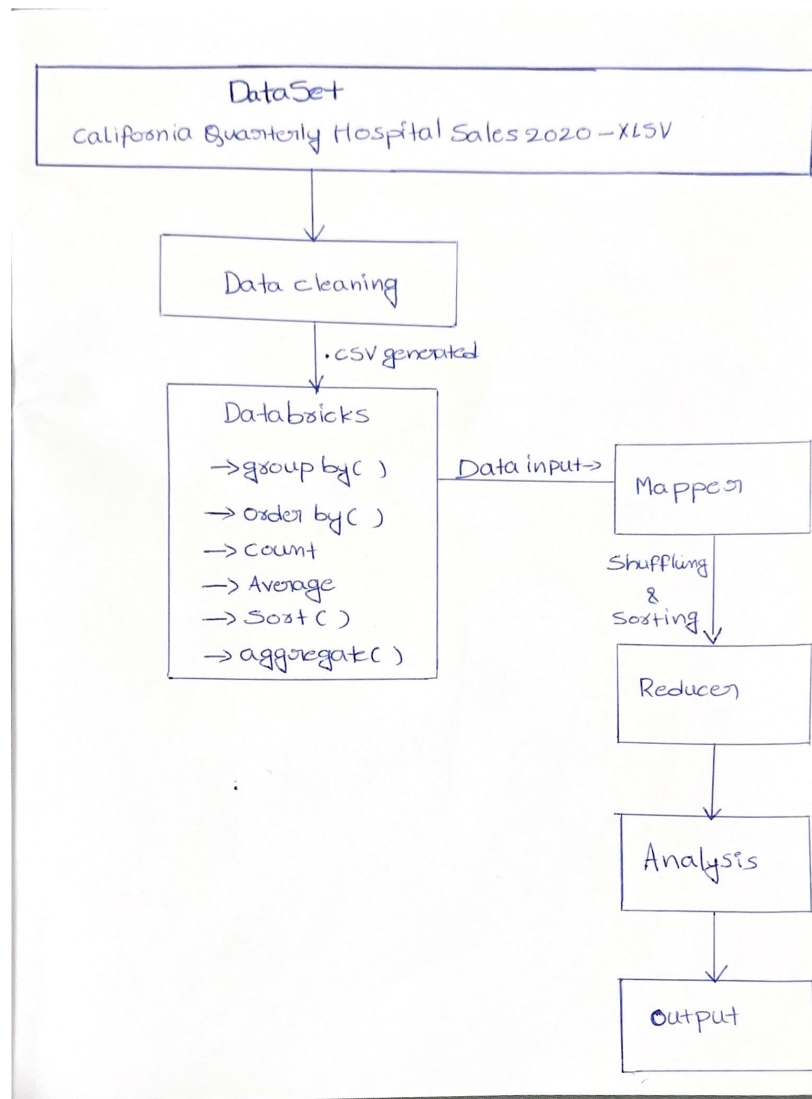


Fig. 1. A detailed methodology in the form of a block diagram

1.5 Detailed explanation of steps of implementation

Implementation of Hadoop MapReduce

1. prerequisite : Eclipse IDE tool
2. Downloaded and Imported the code which already includes the implementation of map Reduce concepts from the course materials.
3. The existing output file need to be deleted by navigating to the project folder MapReduceDemo/data/output.
4. Hadoop consists of a RecordReader that uses TextInputFormat to transform input splits into key-value pairs.
5. The key-value pairs are then used as inputs in the mapping step.
6. The mapper processes the key-value pairs and produces an output of the same form (key-value pairs).
7. Flow of the MAP REDUCE: Input,Splitting,Mapping,Shuffling,Reducing,Final Result
8. Since we would like to work on healthcare,We would be having two goals for which we apply the MapReduce concepts. 4.Segregate the count of hospital list depending upon the central type of each state How many are Non Profit Corp? How many are Church? 8.Categorize the hospital based on it's type How many are Psychiatric Health Facilities? How many are Comparable?
9. To achieve these two goals we have to extract the columns from the dataset named TYPE CNTRL and TYPE HOSP and use this two columns as input file. In the Mapper class , since we have TYPE CNTRL as first column in the input data need to mention row[0] in the below code. context.write(new Text(row[0]), new IntWritable(1)); and similarly TYPE HOSP is the second column so we need to mention row[1] in the below code, context.write(new Text(row[1]), new IntWritable(1));
10. In the mapper class, we split the input dataset into chunks
11. Reducer process these intermediate data from the maps into smaller tuples, that reduces the tasks, leading to the final output of the framework
12. In Reducer class, the count functionality is implemented to get the count of each Type of hospital and type of central.
13. we need to run the code and check the output on project directory MapReduceDemo/data/output , In the output folder we need to open the file with .txt extension in order to see the results.
14. Steps to Follow Databricks
15. Step 1: We must register for a databricks account using the link below. <https://community.cloud.databricks.com/>
16. step 2: After successfully logging in, you may use PySpark and SQL in DataBricks to execute the code. For Goals 1,2,3,5,6,7,9 and 10, we had applied pyspark and SQL commands to obtain the results.
17. For Goal1:In which locality highest no of health care units are observed?
18. To getresults for this goal we can use either pyspark or sql
19. 1. First we need to create a data frame
20. 2. we can display the data by using below code

21. 3. File location and type `file_location = "/FileStore/tables/2020_quarter2_complete_data_set10192020-3.csv"` `file_type = "csv"`
22. 4.CSV options `infer_schema = "false"` `first_row_is_header = "true"` `delimiter = ","`
23. 5. The applied options are for CSV files. For other file types, these will be ignored.
`df = spark.read.format(file_type).option("inferSchema", infer_schema).option("header", first_row_is_header).display(df)`
 Create a view or table
`temp_table_name = "Hospital_Data" df.createOrReplaceTempView(temp_table_name)`
24. step 3: we can now able to see the data by using below SQL command
`/* Query the created temp table in a SQL cell */`
`select * from 'Hospital_Data'`
25. step 4: Run the code below to achieve Goal 1's highest number of healthcare units.
`SELECT COUNTY_NAME, COUNT(COUNTY_NAME) AS no_Hsptls FROM Hospital_Data GROUP BY COUNTY_NAME (SELECT MAX(mycount) FROM (SELECT COUNTY_NAME, COUNT(COUNTY_NAME) AS mycount FROM Hospital_Data))`
26. step 5: Run the code below to achieve Goal 2's Lowest number of healthcare units.
`SELECT COUNTY_NAME, COUNT(COUNTY_NAME) AS no_Hsptls FROM Hospital_Data GROUP BY COUNTY_NAME (SELECT MIN(mycount) FROM (SELECT COUNTY_NAME, COUNT(COUNTY_NAME) AS mycount FROM Hospital_Data))`
27. Step6:Goal3: Run the below code for Goal3, Cities with no.of healthcare centers to achieve the obtained results.
`/* Cities with no.of healthcare centers */ SELECT COUNTY_NAME, COUNT(*) AS no_hsptls FROM Hospital_Data`
28. step7: For Goal5 For which hospital highest no of available beds are recorded?
 ,run the below code to get results.
 For SQL:
`SELECT FAC_NAME, COUNTY_NAME, AVL_BEDS FROM Hospital_Data WHERE AVL_BEDS = (SELECT MAX(AVL_BEDS + 0) FROM Hospital_Data)`
 With PySpark:
`Hospitaldataset = Hosiptaldataset.select("FAC_NAME", "AVL_BEDS").where("AVL_BEDS > 0") Hosiptaldataset = Hosiptaldataset.sort("AVL_BEDS") display(Hosiptaldataset)`
29. Step8:Run the below code for Goal6 For which hospital least no of available beds are recorded?
`/* Health care centers with lowest available beds */ FAC_NAME, COUNTY_NAME, AVL_BEDS FROM Hospital_Data (SELECT MIN(AVL_BEDS + 0) FROM Hospital_Data)`
30. Step9:Run the below code for Goal7: For to Display the count of distinct HSA on overall health care data
`from pyspark.sql.functions import count`
`countevent = hospital_data.select("HSA").sort("HSA").dropDuplicates()`
`display(countevent)`
31. Step10: Run the below code for Goal9 Highest capacity of licenced beds for hospitals
 with SQL: `/* Health care centers with highest licensed beds */ SELECT FAC_NAME, COUNTY_NAME, LIC_BEDS (SELECT MAX(LIC_BEDS + 0) FROM Hospital_Data)`
 With PySpark: HIGHEST CAPACITY OF LICENCED BEDS from pyspark.sql.functions
`import count Hosiptaldataset = Hosiptaldataset.select("CITY", "LIC_BEDS").groupBy("CITY").agg(count("LIC_BEDS"))`
`display(Hosiptaldataset)`

32. Step11:Goal10: Run the below code lowest capacity of licenced beds for hospitals to acheive the the obtained results.

```
/* Health care centers with lowest licensed beds */ SELECT FAC_NAME, COUNTY_NAME, LIC_BEDS FROM
(SELECT MIN(LIC_BEDS + 0) FROM Hospital_Data)
```

33. Using the entire above mentioned code, we can view the outcomes for the stated goals.

1.6 Discussion of results of each goal

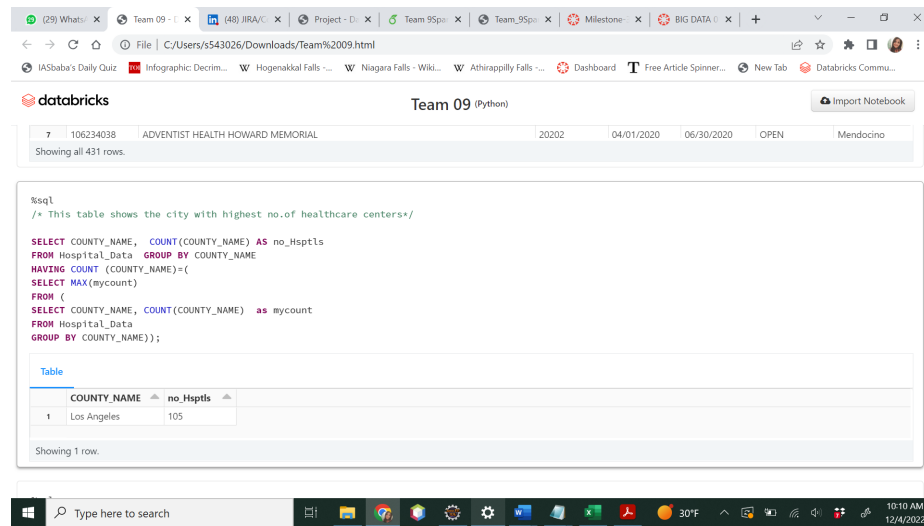


Fig. 2. Highest number of Health Care centers

The output Displayed for the Goal1:

1. The above figure shows results for In which locality highest no of health care centers are observed? According to the results Los angels topped the highest health care centers with 105 health care units in it.The code took about 20 minutes to write and run. and did not face any major errors or problems except login issues with Data bricks

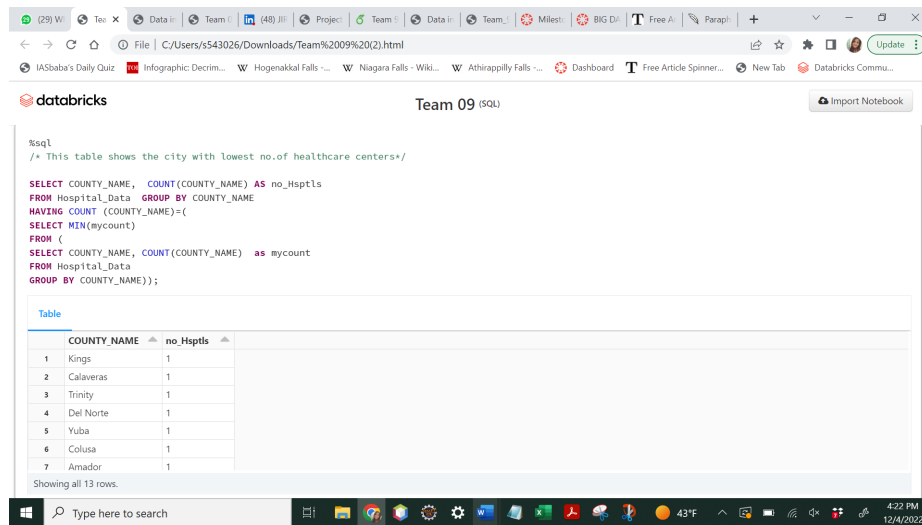
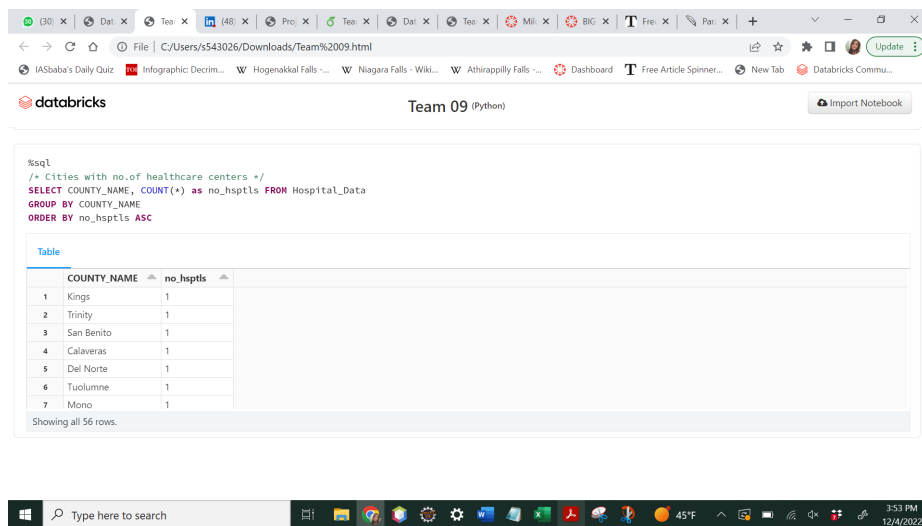


Fig. 3. Lowest number of Health Care centers

The output Displayed for the Goal2:

2. 1. The above figure shows results for In which locality Lowest no of health care centers are observed? According to the results there are multiple cities with Lowest health care centers .The code took about 15 minutes to write and run. and did not face any major errors or problems except login issues with Data bricks, People who desire to live in California or those who are new to the state can quickly determine which locality is suited for their living with the help of these results.



The screenshot shows a Databricks notebook interface. At the top, there's a browser window with multiple tabs. Below that, the Databricks logo and 'Team 09 (Python)' are visible. The main area contains an SQL query and its results.

```
%sql
/* Cities with no.of healthcare centers */
SELECT COUNTY_NAME, COUNT(*) as no_hspts FROM Hospital_Data
GROUP BY COUNTY_NAME
ORDER BY no_hspts ASC
```

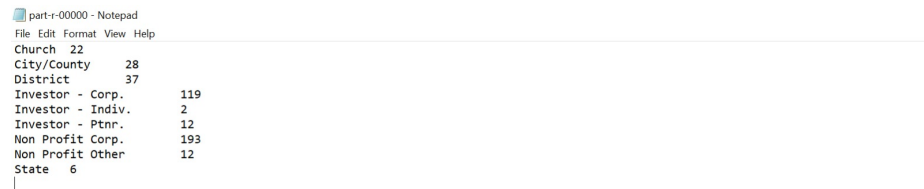
	COUNTY_NAME	no_hspts
1	Kings	1
2	Trinity	1
3	San Benito	1
4	Calaveras	1
5	Del Norte	1
6	Tuolumne	1
7	Mono	1

Showing all 56 rows.

Fig. 4. Cities with no.of healthcare centers

Output Displayed for the Goal 03:

- 1.The above figure shows results of number of healthcare centers in cities According to the results, we can clearly see the cities and number of healthcare centers, to get these results i have used SQL commands by using data bricks it took 30 minutes to write and execute code and i did not face any major problems while writing and executing the code



```

part-r-00000 - Notepad
File Edit Format View Help
Church 22
City/County 28
District 37
Investor - Corp. 119
Investor - Indiv. 2
Investor - Ptnr. 12
Non Profit Corp. 193
Non Profit Other 12
State 6
|

```

Fig. 5. Output Showing the Count of each Central Type

The Output displayed for the goal4: 4.Segregate the count of hospital list depending upon the central type of each state How many are Non Profit Corp? How many are Church? The above output shows the Count of each individual type of Central.From the Output we can see highest number of Non Profit Corp. and Lowest number of Investor- Individual It took around 20 minutes to execute the code and Did not face any difficulties while executing the code.

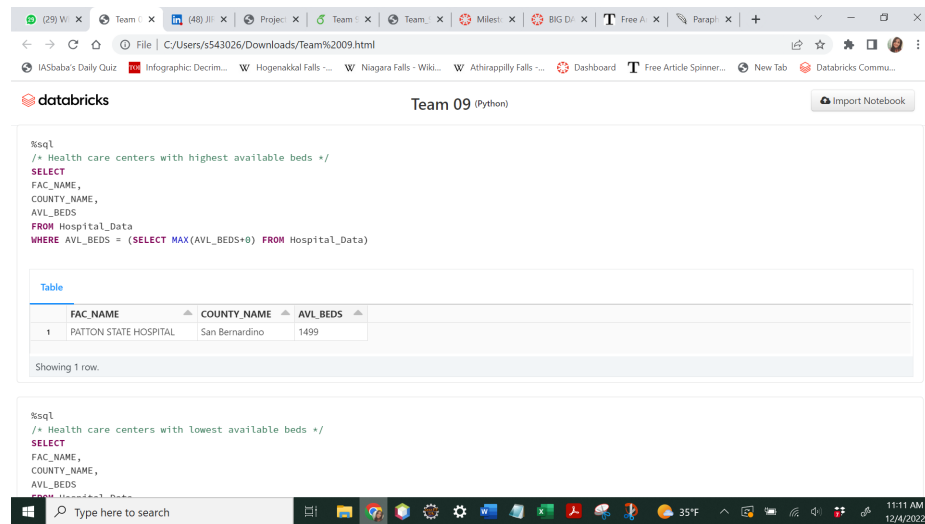


Fig. 6. hospitals with highest no of available beds are recorded

Output Displayed for the Goal5:

4. 1.The above figure shows results For which hospital highest no of available beds are recorded According to the results PATTON STATE HOSPITAL recorded top with 1499 available beds and it is located in San Bernardino, to get these results i have used SQL commands by using data bricks it took 20 minutes to write and execute code and i did not face any major problems while writing and executing the code

The screenshot shows a Databricks notebook interface. The top bar indicates the workspace is 'Team 09 (Python)'. The notebook contains two SQL queries. The first query, titled 'Health care centers with lowest available beds', uses a subquery to find the minimum available beds across all hospitals and then filters for the hospital with that minimum value. The output is a table with one row: PATIENTS' HOSPITAL OF REDDING, located in Shasta, with 10 available beds. The second query, titled 'Health care centers with highest licensed beds', is partially visible but its output is not shown. The bottom of the screen shows a Windows taskbar with the date 12/4/2022 and time 11:21 AM.

```
%sql
/* Health care centers with lowest available beds */
SELECT
  FAC_NAME,
  COUNTY_NAME,
  AVL_BEDS
FROM Hospital_Data
WHERE AVL_BEDS = (SELECT MIN(AVL_BEDS+0) FROM Hospital_Data)
```

	FAC_NAME	COUNTY_NAME	AVL_BEDS
1	PATIENTS' HOSPITAL OF REDDING	Shasta	10

Showing 1 row.

```
%sql
/* Health care centers with highest licensed beds */
SELECT
  FAC_NAME,
  COUNTY_NAME,
  LIC_BEDS
FROM Hospital_Data
```

Fig. 7. hospitals with Lowest no of available beds are recorded
Output Displayed for Goal6:

5. 1.The above figure shows results For which hospital lowest no of available beds are recorded According to the results PATIENTS' HOSPITAL OF REDDING recorded lowest available beds with 10 beds and it is located in Shasta, to get these results i have used SQL commands by using data bricks it took 20 minutes to write and execute code and i did not face any major problems while writing and executing the code

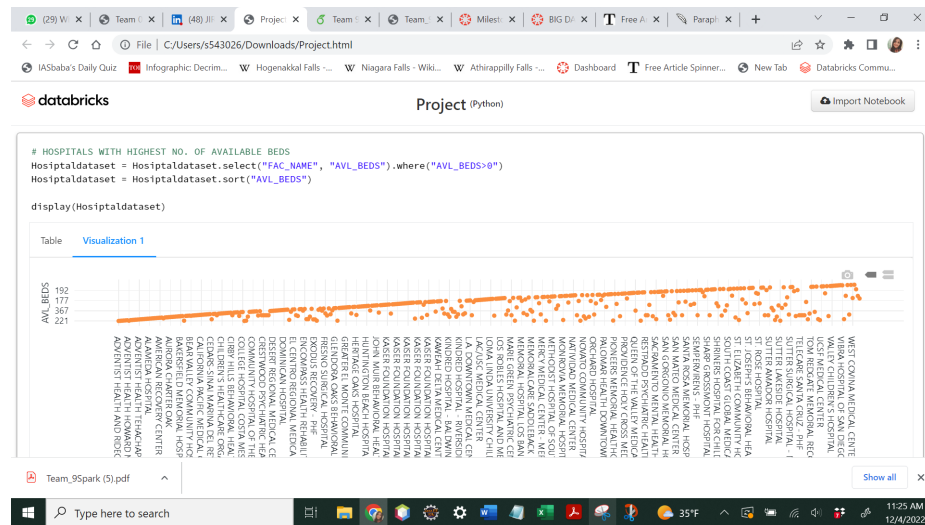


Fig. 8. hospitals with highest and Lowest no of available beds are recorded
Output Displayed for the Goal 5 and 6 using Pyspark:

6. 1.The above figure shows results For which hospital highest and lowest no of available beds are recorded According to the results PATTON STATE HOSPITAL, PATIENTS' and HOSPITAL OF REDDING recorded highest and lowest available beds with 1499 and 10 beds and located in San Bernardino and Shasta, to get these results i have used pyspark commands by using data bricks it took 30 minutes to write and execute code and i did not face any major problems while writing and executing the code

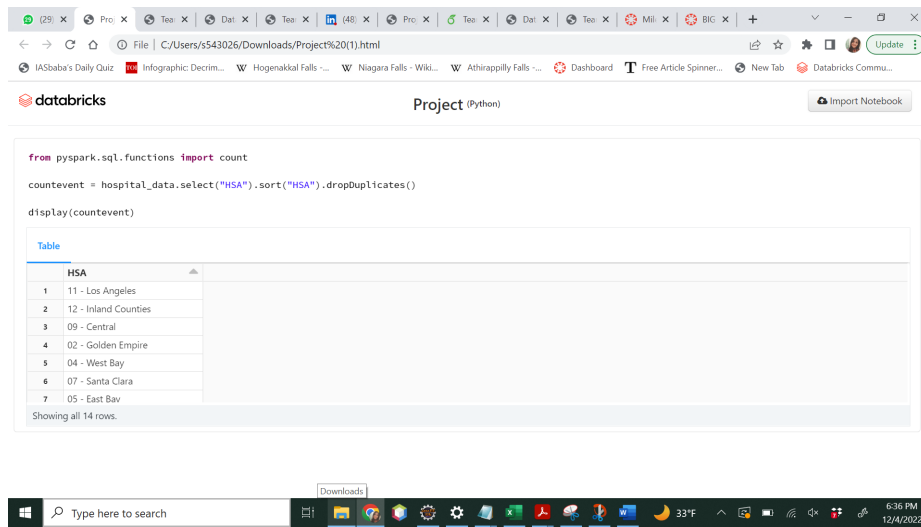


Fig. 9. Display the count of distinct HSA on overall health care data

The Output displayed for the goal7: Find the count of hospitals in a locality who has HSA(Health Savings Account) Output Displayed for the below Goal: The above figure shows results of Health Service Area (HSA) of hospitals According to the results over all there are 14 different HSA and are present among 431 hospitals and in those 14 there are repeated several times i.e The hospitals have same HSA to get these output using a pyspark command dropDuplicates which filter and display the exact number of HSA, it took 30 minutes to write and excute the code, did not face any major problems.



```
part-00000 - Notepad
File Edit Format View Help
Comparable      359
Hospital-LTC Emphasis  4
Kaiser Foundation Health  31
Psychiatric Health Facilities  30
Shriners Hospitals      1
State Hospitals  6
|
```

Fig. 10. Output Showing the Count of each Type of the Hospital

The Output displayed for the goal8: Goal8.Categorize the hospital based on it's type
How many are Psychiatric Health Facilities? How many are Comparable?

The above output shows the Count of each individual type of Hospital.From the Output
we can see highest number of Comparable Hospital and Lowest number of shiners
Hospital. It took around 20 minutes to execute the code and Did not face any difficulties
while executing the code.

The screenshot shows a Databricks notebook interface. The top bar indicates the notebook is for 'Team 09 (Python)'. The main area contains a SQL query that selects the hospital name, county name, and licensed beds from the 'Hospital_Data' table, where the licensed beds are equal to the maximum licensed beds in the same table. The query is executed, and the output is displayed as a table with one row: COALINGA STATE HOSPITAL in Fresno with 1500 licensed beds. Below the first query, a second SQL query is visible, which selects the hospital name and county name for the lowest licensed beds.

```
%sql
/* Health care centers with highest licensed beds */
SELECT
  FAC_NAME,
  COUNTY_NAME,
  LIC_BEDS
FROM Hospital_Data
WHERE LIC_BEDS = (SELECT MAX(LIC_BEDS+0) FROM Hospital_Data)
```

	FAC_NAME	COUNTY_NAME	LIC_BEDS
1	COALINGA STATE HOSPITAL	Fresno	1500

Showing 1 row.

```
%sql
/* Health care centers with lowest licensed beds */
SELECT
  FAC_NAME,
  COUNTY_NAME
```

Fig. 11. Highest capacity of licensed beds for hospitals

Output Displayed for the Goal 9:

7. 1. The above figure shows results Highest capacity of licensed beds for hospitals. According to the results COALINGA STATE HOSPITAL recorded highest licensed beds located in Fresno with 1500 beds, to get these results I have used SQL commands by using data bricks it took 30 minutes to write and execute code and I did not face any major problems while writing and executing the code.

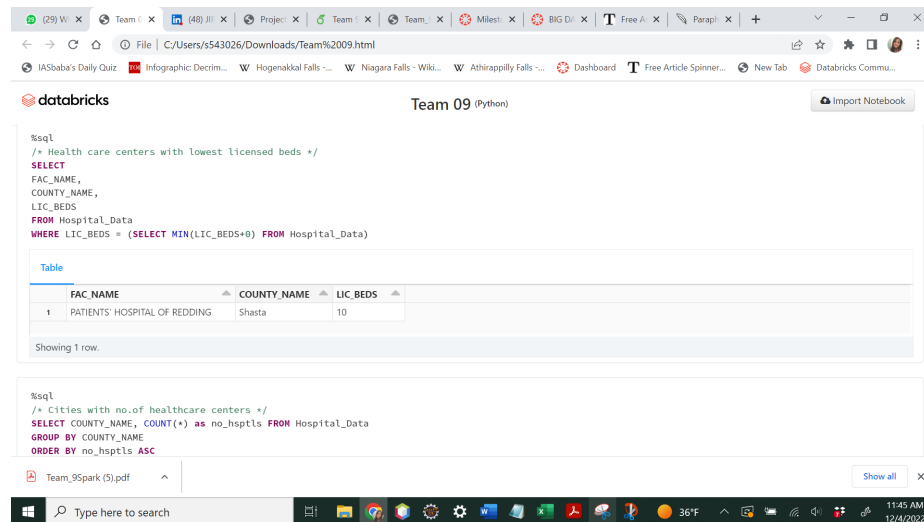


Fig. 12. Lowest capacity of licensed beds for hospitals

Output Displayed for the Goal10:

8. 1. The above figure shows results Lowest capacity of licensed beds for hospitals According to the results PATIENTS' HOSPITAL OF REDDINGL recorded lowest licensed beds located in Shasta with 10 beds, to get these results i have used SQL commands by using data bricks it took 30 minutes to write and execute code and i did not face any major problems while writing and executing the code

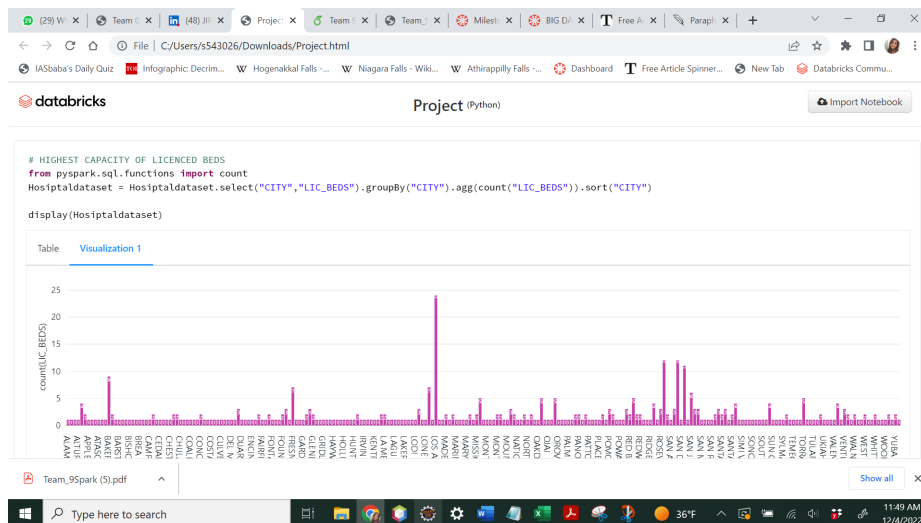


Fig. 13. Lowest capacity of licensed beds for hospitals

Output Displayed for the Goal 9 and 10 using Pyspark:

9. 1.The above figure shows results highest and Lowest capacity of licensed beds for hospitals

According to the results, COALINGA STATE HOSPITAL,PATIENTS' HOSPITAL OF REDDINGL recorded lowest licensed beds located in Fresnoand Shasta with 1500 and 10 beds, to get these results i have used Pyspark commands by using data bricks it took 30 minutes to write and execute code and i did not face any major problems while writing and executing the code

10. CONCLUSION:

11. 1.We successfully accomplished our objectives with the help of Hadoop, MapReduce, DataBricks, SQL, and PySpark. The primary goal of analysis is to provide useful information for those who are looking for healthcare information,
12. 2.From the above results we can conclude that highest number of healthcare centers are in Los angels and lowest number of number of health care centers are observed in multiple cities from this we can say that more health care centers need to be built in that respective areas.
13. 3.Based on the type of central there are highest number of hospitals are categorized as non profit corporations, and lowest number of hospitals are categorized as investor-individual.
14. 4.Patton state hospital recorded highest availble bed in san barnardino, from this we can conclude people near by in this locality can prefer this hospital who are looking for available beds
15. 5.patients hospital of redding recorded lowest available beds located in shata
16. GitHub Link: