COLLEGE CODE: 5113

Batch Members:

1.MOHAMMED USAID T -(au511321104054)- mohammedusaidt@gmail.com
2. LINGESHWARAN D -(au511321104048)- **lingesh252004@gmail.com**
3.GUNASEELAN J -(au511321104026)- jsgunaseelan2004@gmail.com

CLOUD APPLICATION DEVELOPMENT

PROJECT 5:BIG DATA ANALYSIS WITH IBM CLOUD DATABASE

Navigating the Path to Big Data Insights with IBM Cloud Databases

1.1. Introduction

In today's data-driven world, the ability to harness the power of big data is essential for informed decision-making and innovation. The "Big Data Analysis" project is a comprehensive exploration of data analytics using IBM Cloud Databases. It seeks to extract valuable insights from extensive datasets, including climate trends and social patterns.

1.2. Problem Statement

The primary challenge is to delve into the world of big data analysis using IBM Cloud Databases. The goal is to uncover hidden insights within these datasets, which may include data related to climate trends or social media patterns. The project also includes designing the analysis process, setting up IBM Cloud Databases, conducting data analysis, and creating visualizations to derive essential business intelligence.

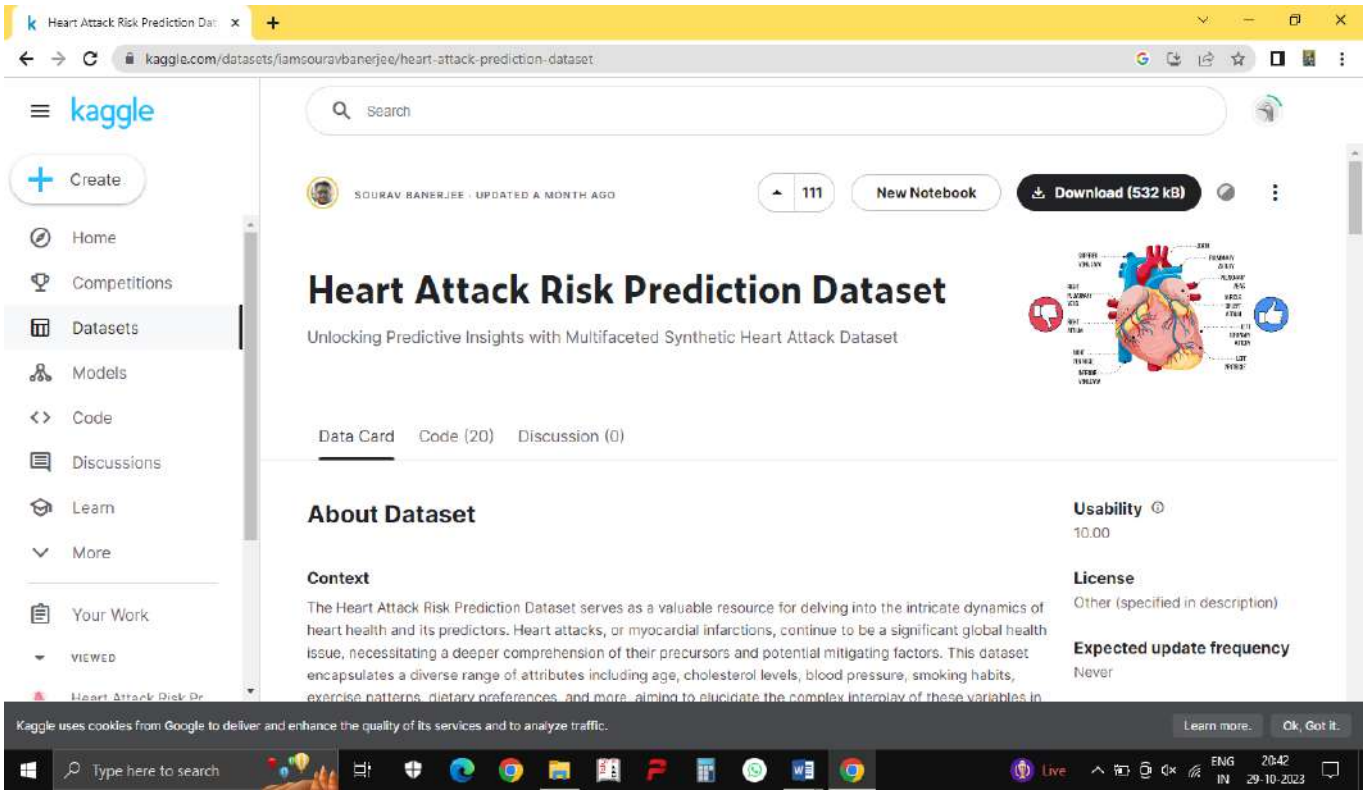
1.3. Objectives

- `Identify and select relevant datasets for analysis.
- Configure IBM Cloud Databases for efficient data storage and management.
- Develop queries and scripts for data exploration.
- Apply appropriate analysis techniques, such as statistical analysis and machine learning, to extract insights.
- Create effective visualizations to present the analysis results.
- Interpret the findings to derive actionable business recommendations.

2.Understanding the Problem Statement

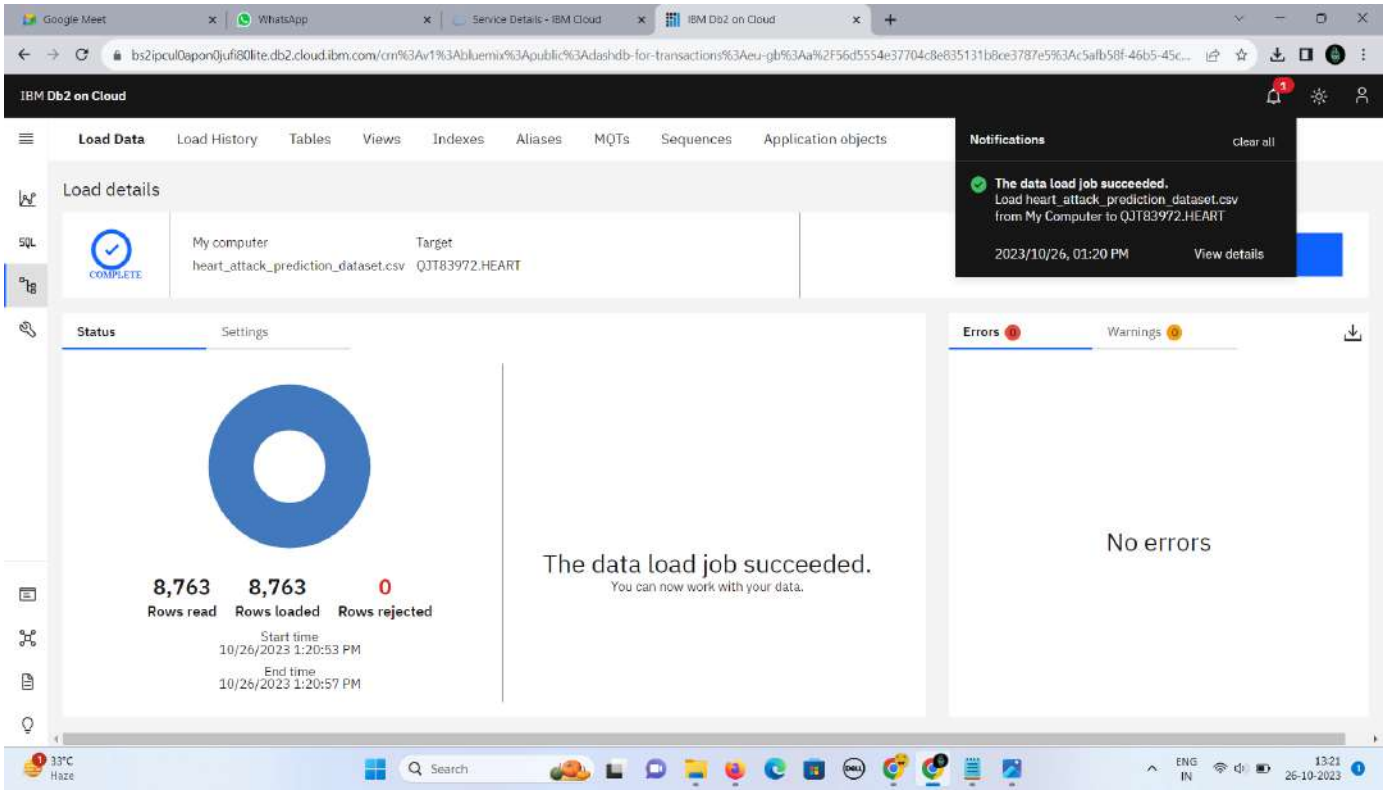
2.1. Data Selection

To address the problem statement effectively, the first step involves selecting appropriate datasets. These datasets should align with the project's objectives and can include climate data, social media trends, or other relevant data sources. The key challenge here is to ensure that the selected datasets are comprehensive and contain valuable information



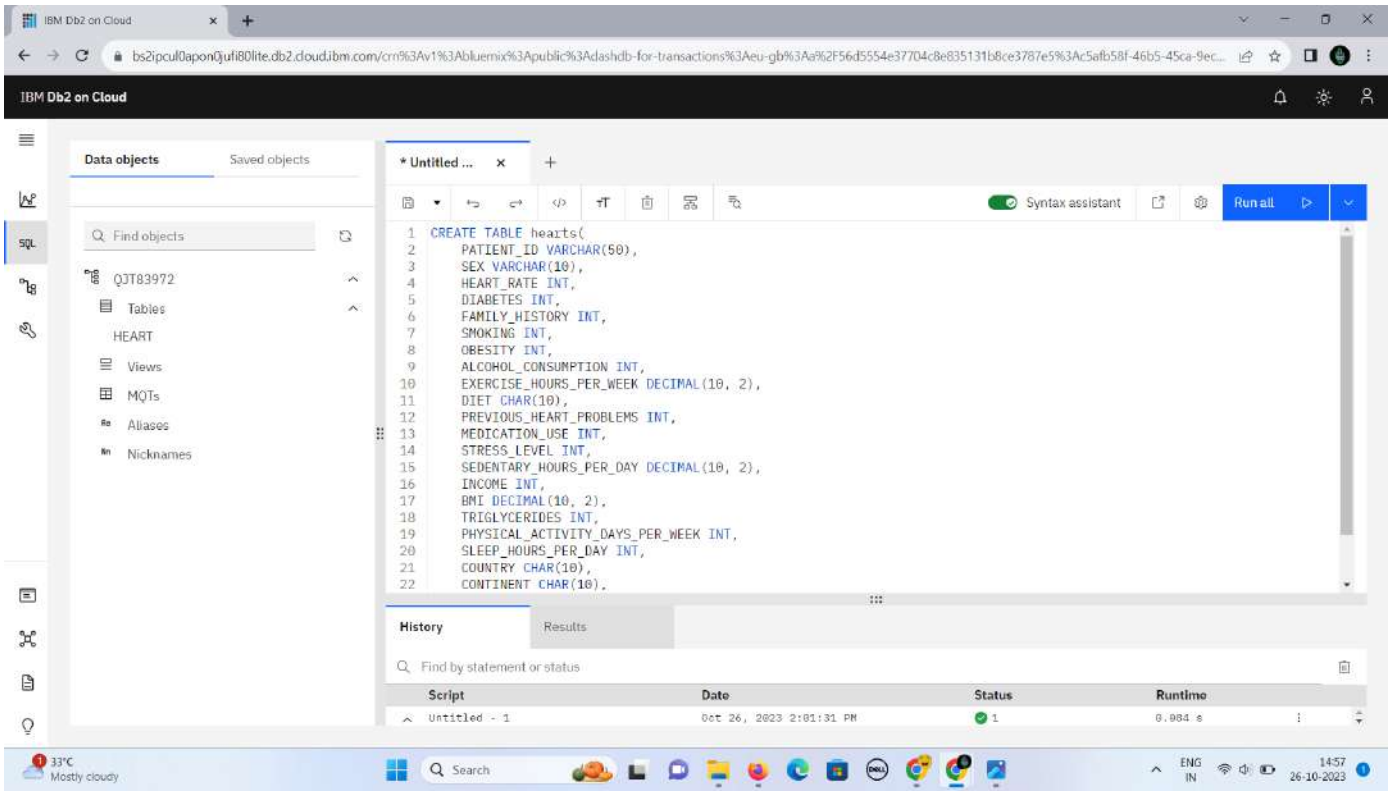
2.2. Database Setup

Efficient data management is pivotal in big data analysis. We will set up IBM Cloud Databases to store and manage the selected datasets. This step demands a keen understanding of the database infrastructure and proper configuration to handle large volumes of data effectively



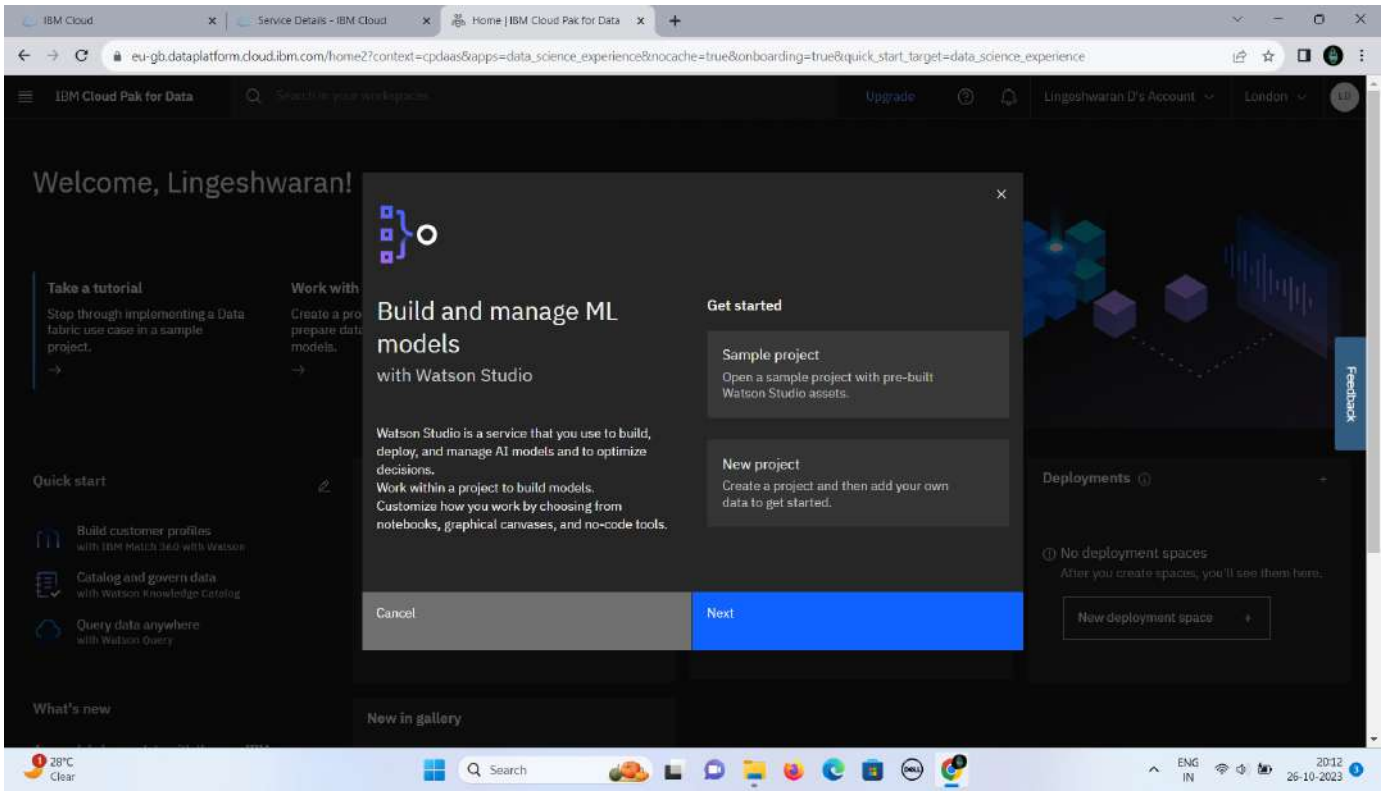
2.3. Data Exploration

The heart of this project lies in data exploration. This phase involves developing queries and scripts to delve into the datasets, extract relevant information, and identify patterns. It requires a thorough understanding of the datasets and an ability to navigate through extensive data efficiently.



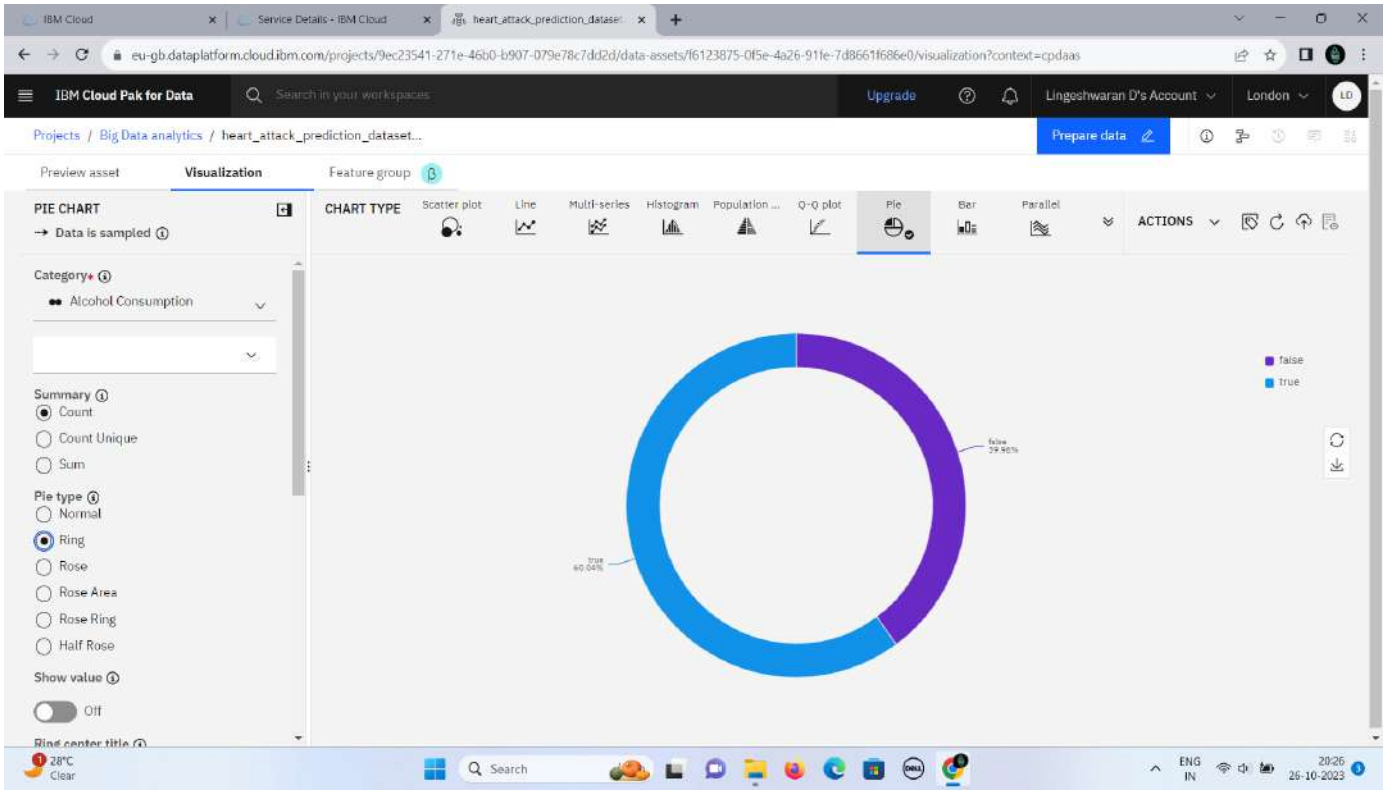
2.4. Analysis Techniques

The chosen datasets may vary in complexity, and different analysis techniques may be needed to extract insights. We will apply appropriate techniques, such as statistical analysis or machine learning, depending on the nature of the data. It's crucial to have a strong grasp of these techniques to ensure meaningful results.



2.5. Visualization

Effective communication of results is essential. To this end, we will design visualizations that not only present the analysis findings but also make them understandable and impactful. This phase demands creative and data visualization skills.



2.6. Business Insights

Business insights derived from big data analytics can provide organizations with valuable information and opportunities for informed decision-making. Here are some key insights and benefits that businesses can gain from big data analytics:

1. Customer Insights:

* Customer behavior analysis:

Big data analytics can help businesses understand customer preferences, buying patterns, and trends, allowing for more personalized marketing and product offerings.

*Customer segmentation:

Businesses can identify different customer segments based on demographics, behavior, and preferences, allowing for targeted marketing and product development.

* Churn prediction:

Analytics can help predict when customers are likely to leave, enabling businesses to take proactive measures to retain them.

2. Operational Efficiency:

*Process optimization:

Big data analytics can identify inefficiencies in business processes and suggest improvements to reduce costs and increase productivity.

*Supply chain optimization:

Analyzing data can help optimize the supply chain, reduce inventory costs, and improve order fulfillment.

3. Product Development:

*Market research:

Big data can provide insights into market trends and consumer needs, helping businesses create products and services that are in demand.

*Product quality and performance:

Analyzing customer feedback and product data can help improve product quality and performance.

4. Fraud Detection:

Anomaly detection:

Big data analytics can detect unusual patterns and anomalies in transactions, helping identify and prevent fraudulent activities.

5. Predictive Maintenance:

Equipment health monitoring:

By analyzing data from sensors and IoT devices, businesses can predict when equipment is likely to fail, reducing downtime and maintenance costs.

6. Financial Analysis:

Risk assessment:

Big data analytics can assess financial risks by analyzing data from various sources, helping organizations make informed investment decisions.

Credit scoring:

Businesses can use big data to improve credit scoring models, leading to better lending decisions.

7. Marketing and Advertising:

*Campaign effectiveness:

Analyzing marketing campaigns can help businesses understand which strategies are most effective and optimize their marketing budgets.

*Social media sentiment analysis:

Monitoring social media can provide insights into customer sentiment and public opinion.

8. Competitive Analysis:

*Market positioning:

Businesses can use big data to understand their position in the market and identify opportunities to gain a competitive edge.

*Competitor analysis:

Analyzing data on competitors can help organizations make informed strategic decisions.

9. Human Resources:

*Employee engagement:

Analytics can be used to measure and improve employee engagement and retention.

*Recruitment and talent management:

Big data can help identify the best candidates for job openings and assess employee performance.

10. Risk Management:

*Identifying and mitigating risks: Big data analytics can help organizations identify potential risks, assess their impact, and develop strategies for risk mitigation.

3. Development Phases :

The project is divided into two primary development phases.

3.1. Development Part 1 :

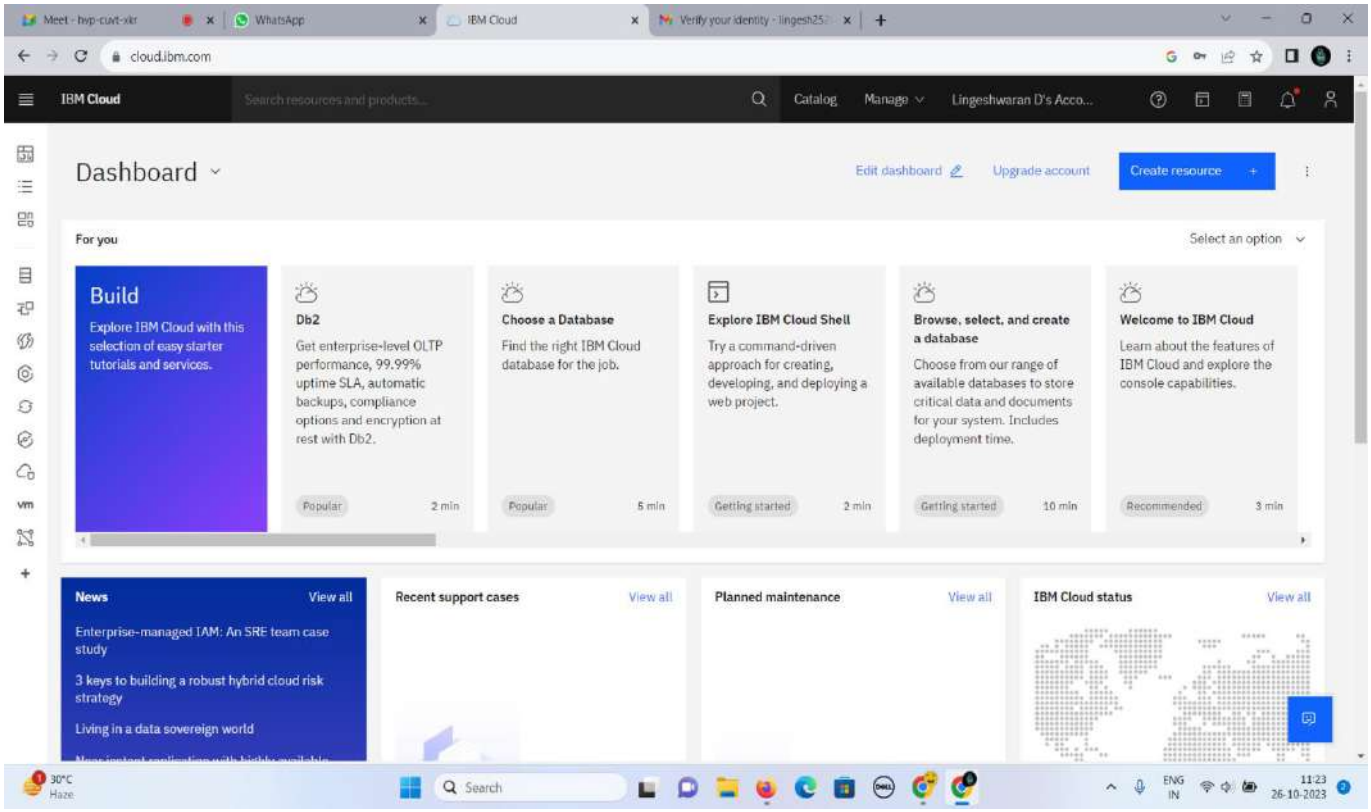
In this phase, we will initiate the big data analysis solution using IBM Cloud Databases. We will import the selected datasets, execute queries and scripts, and start to unveil initial insights.

This part contain’s several steps.

Step 1:

*Login our IBM Cloud Account

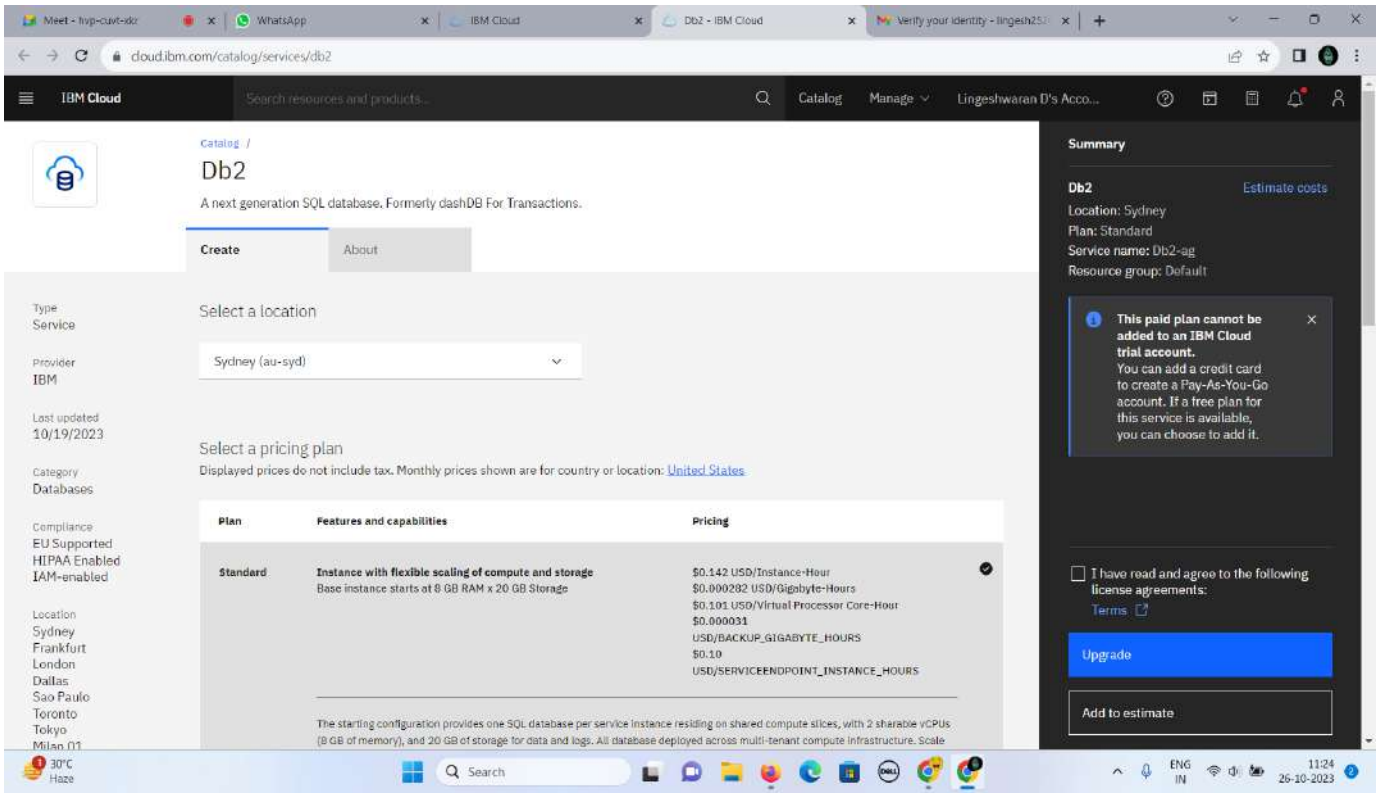
*Search Db2



Step 2:

*Create our free Database 2 server

*Select a location



Step 3:

*Select a location london(eu-gb)

*Create our Database 2 server

Figure 1:

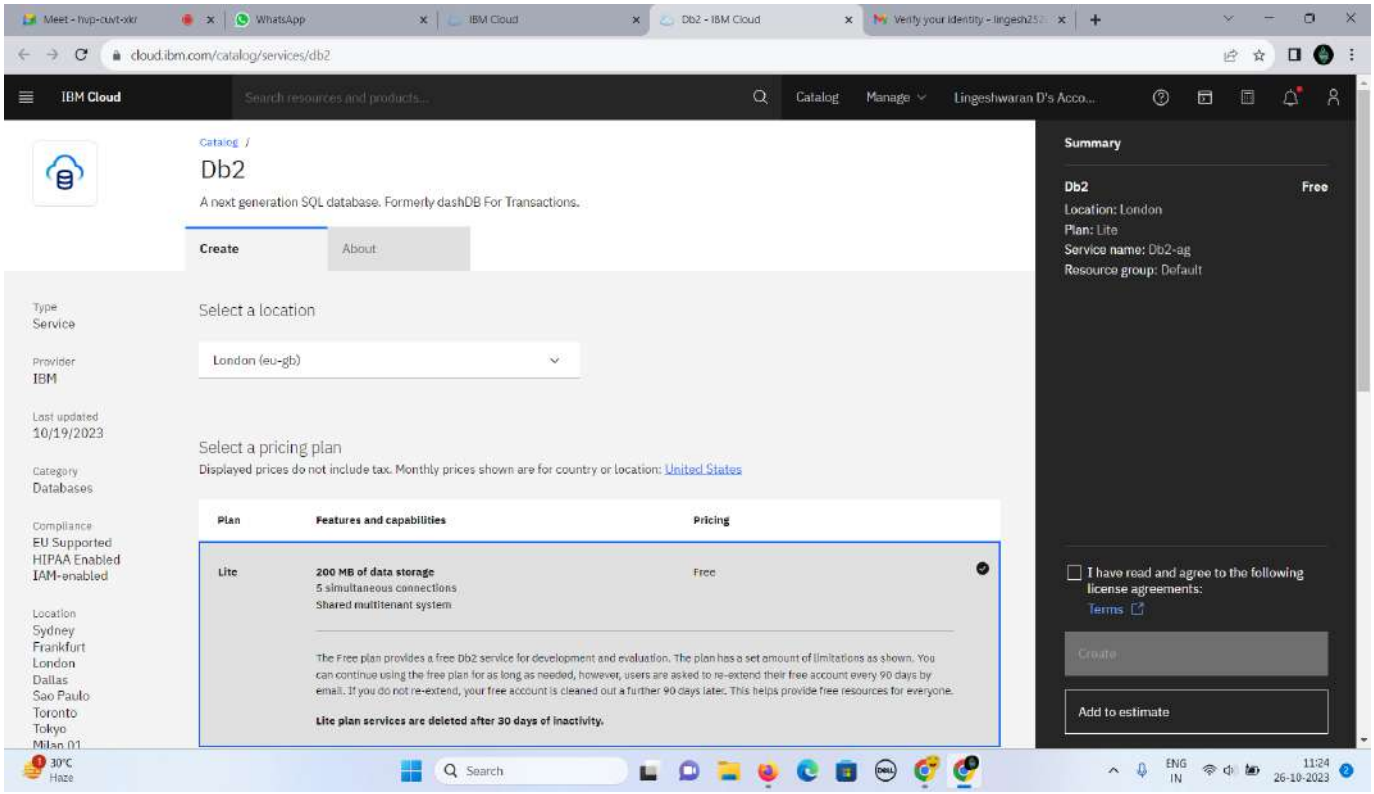
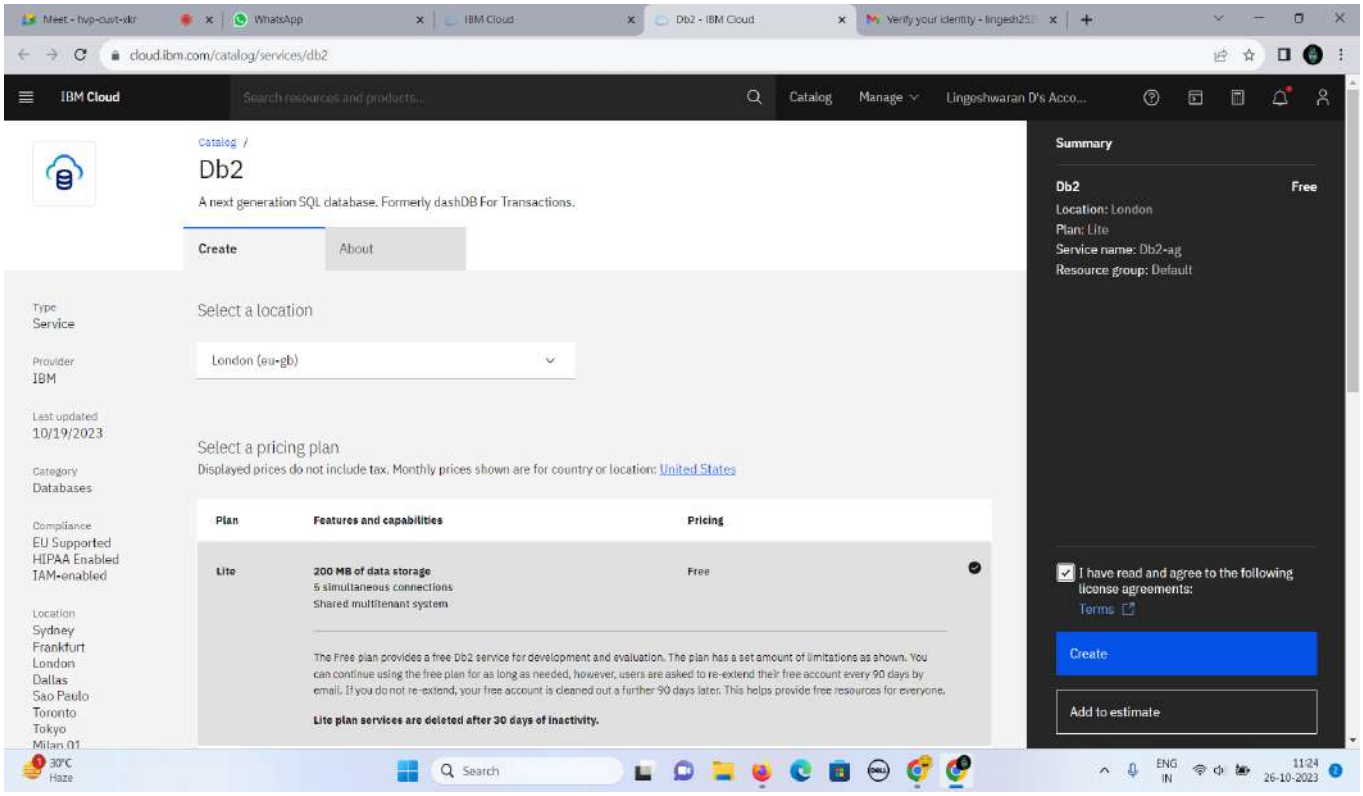


Figure 2:


cloud.ibm.com/catalog/services/db2
IBM Cloud
Search resources and products...
Catalog
Manage
Lingeshwaran D's Acco...
Catalog /
Db2
A next generation SQL database. Formerly dashDB For Transactions.
Create
About
Type
Service
Provider
IBM
Last updated
10/19/2023
Category
Databases
Compliance
EU Supported
HIPAA Enabled
IAM-enabled
Location
Sydney
Frankfurt
London
Dallas
Sao Paulo
Toronto
Tokyo
Milan 01
Select a location
London (eu-gb)
Select a pricing plan
Displayed prices do not include tax. Monthly prices shown are for country or location: United States
Plan
Features and capabilities
Pricing
Lite
200 MB of data storage
5 simultaneous connections
Shared multitenant system
Free
The Free plan provides a free Db2 service for development and evaluation. The plan has a set amount of limitations as shown. You can continue using the free plan for as long as needed, however, users are asked to re-extend their free account every 90 days by email. If you do not re-extend, your free account is cleaned out a further 90 days later. This helps provide free resources for everyone.
Lite plan services are deleted after 30 days of inactivity.
Summary
Db2
Free
Location: London
Plan: Lite
Service name: Db2-ag
Resource group: Default
I have read and agree to the following license agreements:
Terms
Create
Add to estimate
30°C
Haze
Search
ENG
IN
11:24
26-10-2023

Step 4:

*Goto dashboard

*select your Database server 2(Db2-ag)

Figure 1:

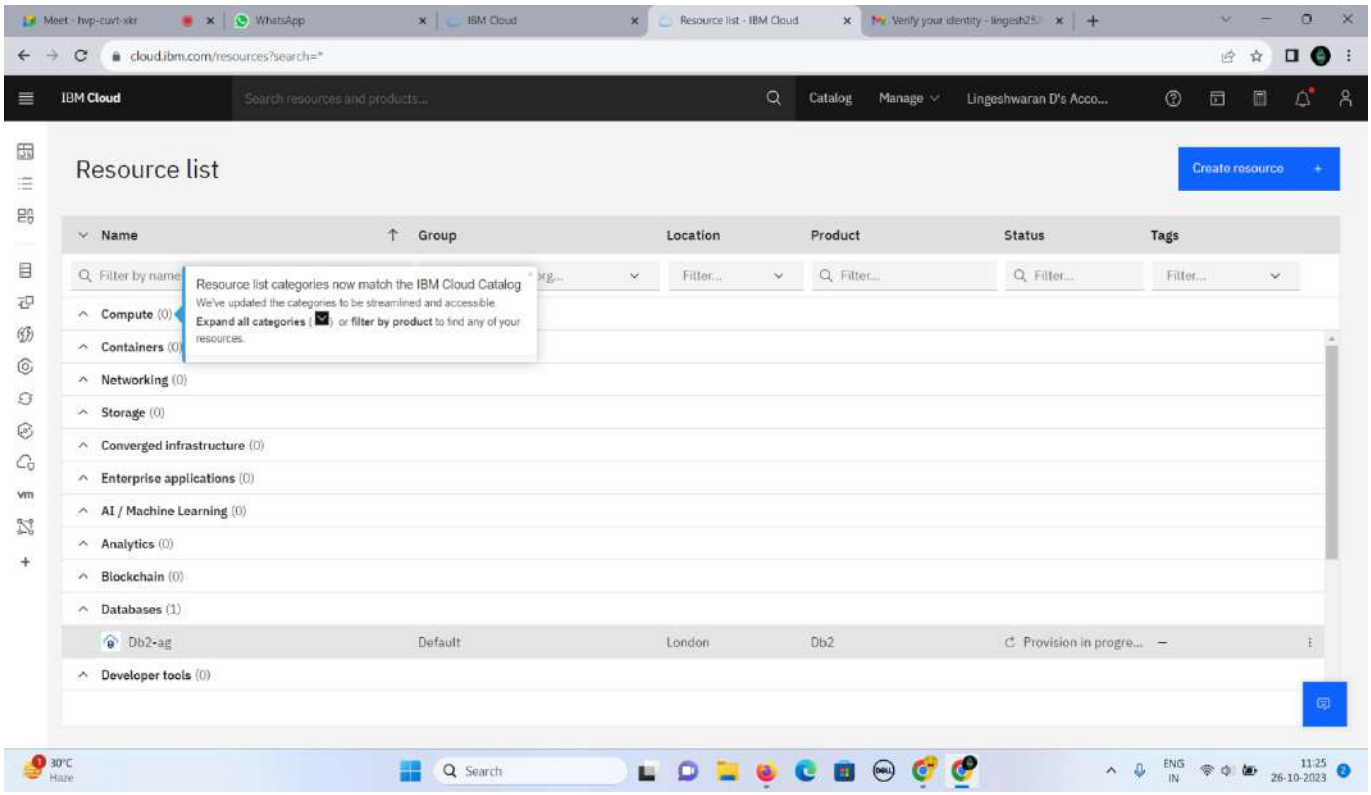
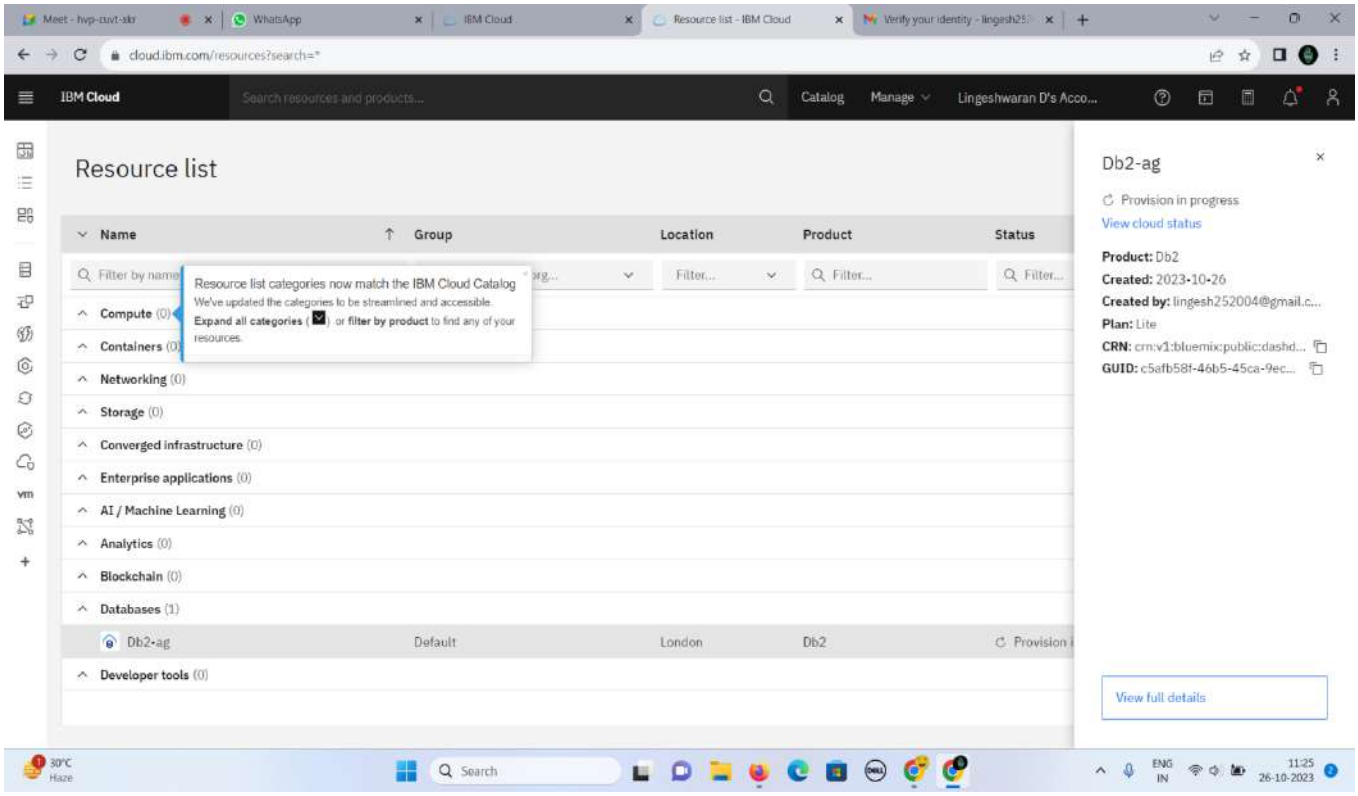
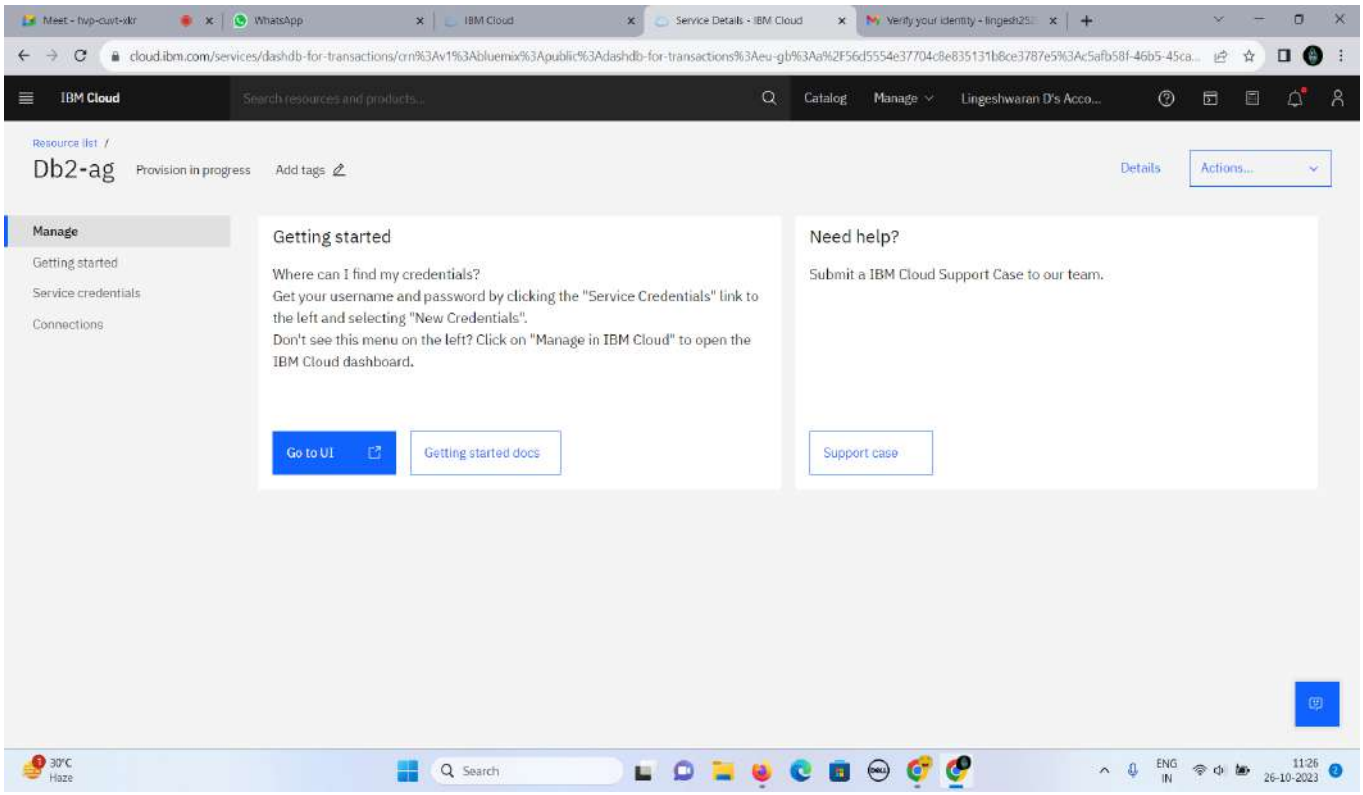


Figure 2:

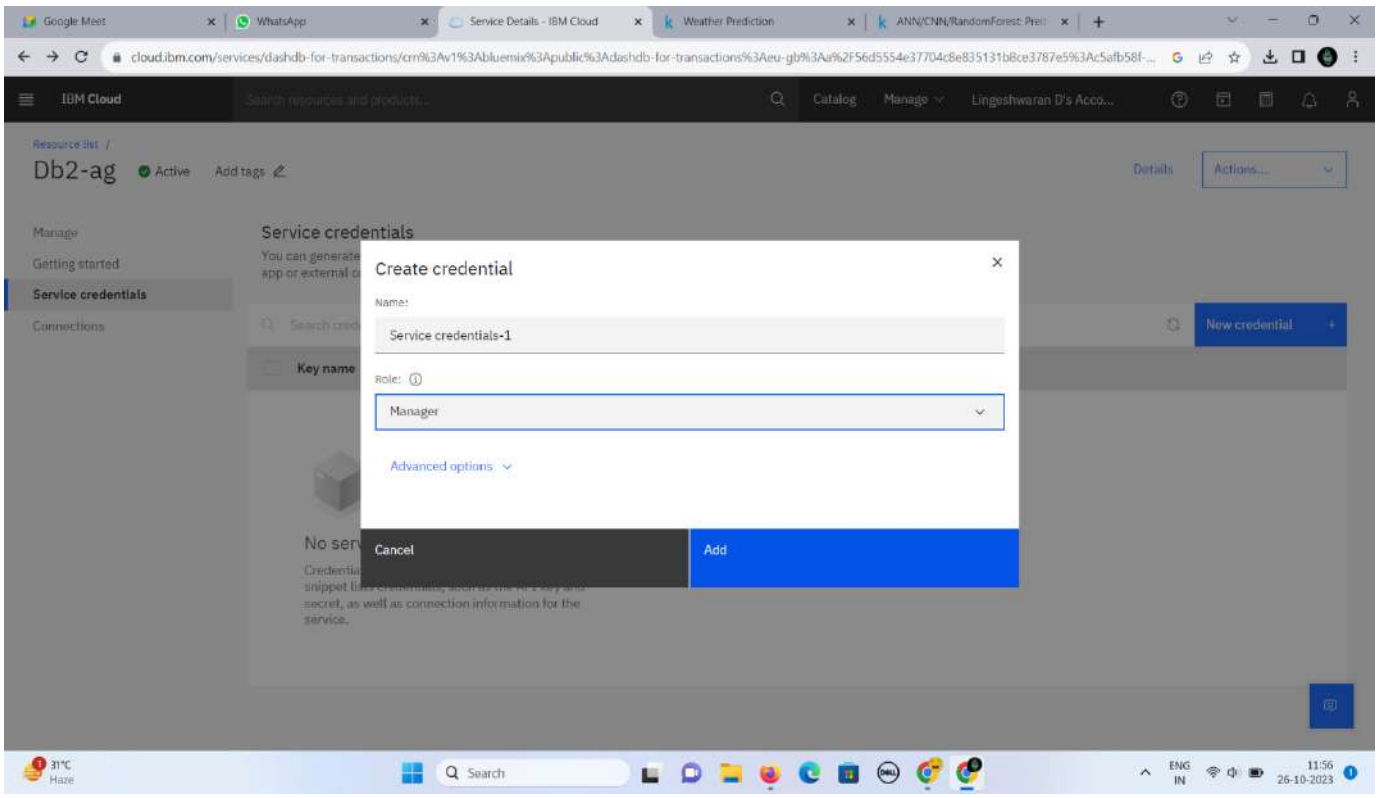

IBM Cloud
Search resources and products...
Catalog
Manage
Lingeshwaran D's Acco...
Resource list
Name
Group
Location
Product
Status
Filter by name...
Filter...
Resource list categories now match the IBM Cloud Catalog
We've updated the categories to be streamlined and accessible.
Expand all categories ( ) or filter by product to find any of your resources.
Compute (0)
Containers (0)
Networking (0)
Storage (0)
Converged infrastructure (0)
Enterprise applications (0)
AI / Machine Learning (0)
Analytics (0)
Blockchain (0)
Databases (1)
Db2-ag
Default
London
Db2
Developer tools (0)
Db2-ag
Provision in progress
View cloud status
Product: Db2
Created: 2023-10-26
Created by: lingesh252004@gmail.c...
Plan: Lite
CRN: crn:v1:bluemix:public:dashd...
GUID: c5afb58f-46b5-45ca-9ec...
View full details
Search
ENG
IN
11:25
26-10-2023

Db2-ag Interface in IBM Cloud

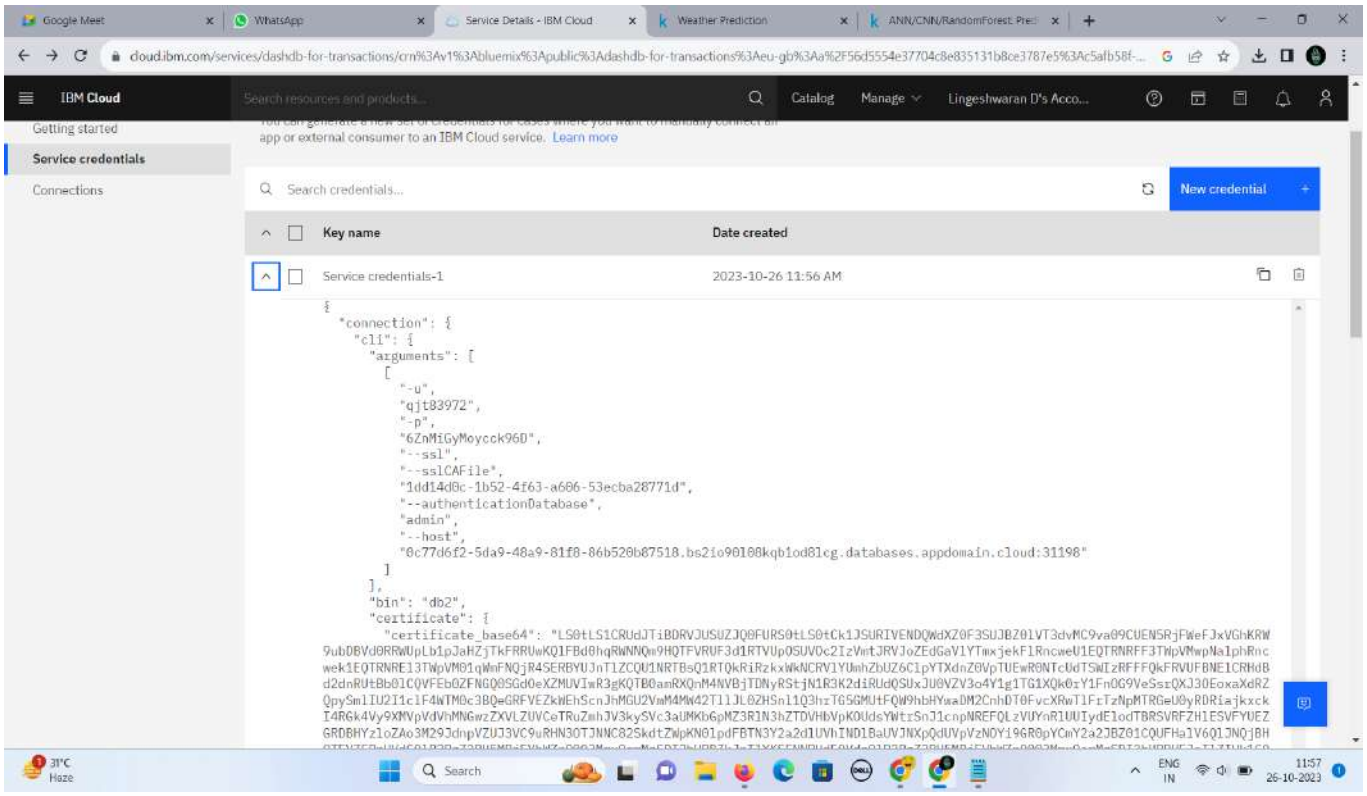


Step 5:

*Create your Server Credential-1

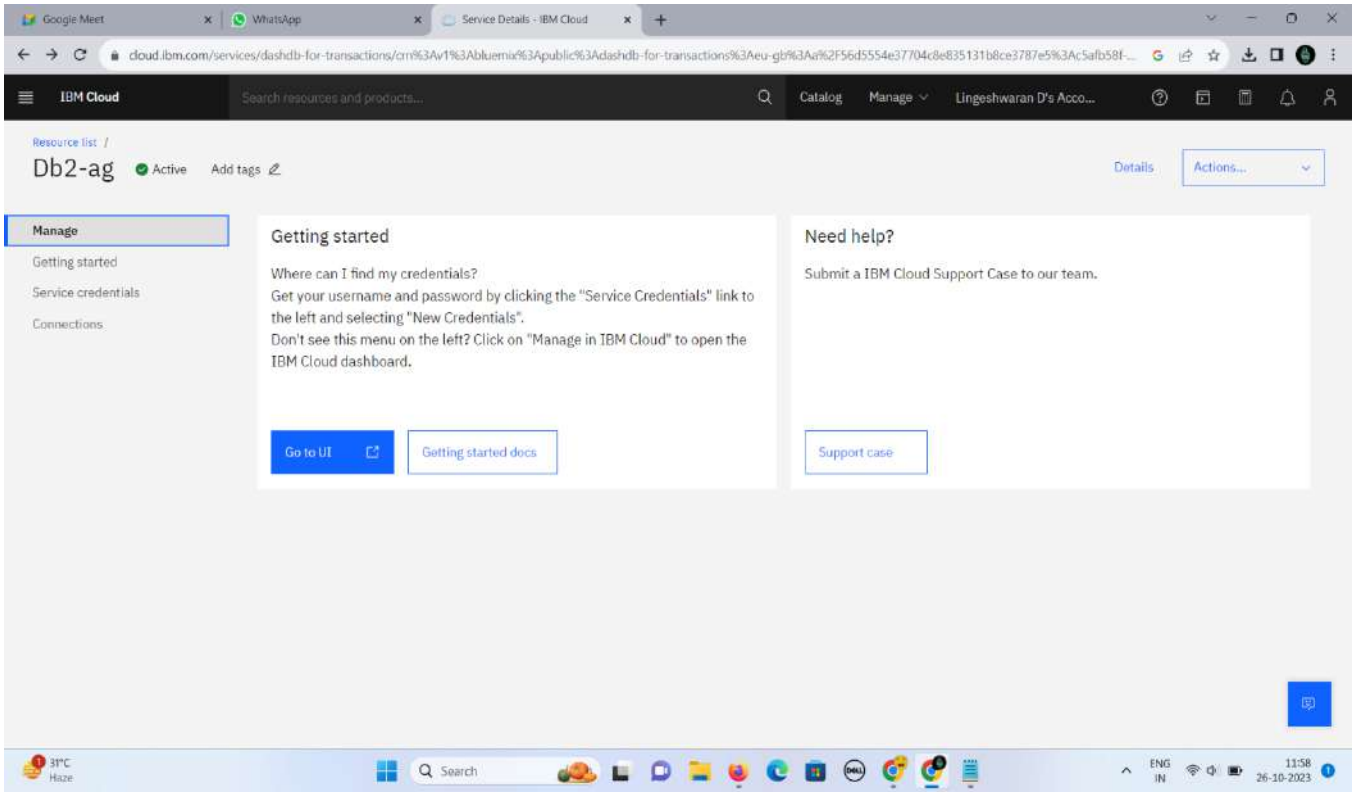


*After create server credential,then it show’s our UserID,Password,Host Address

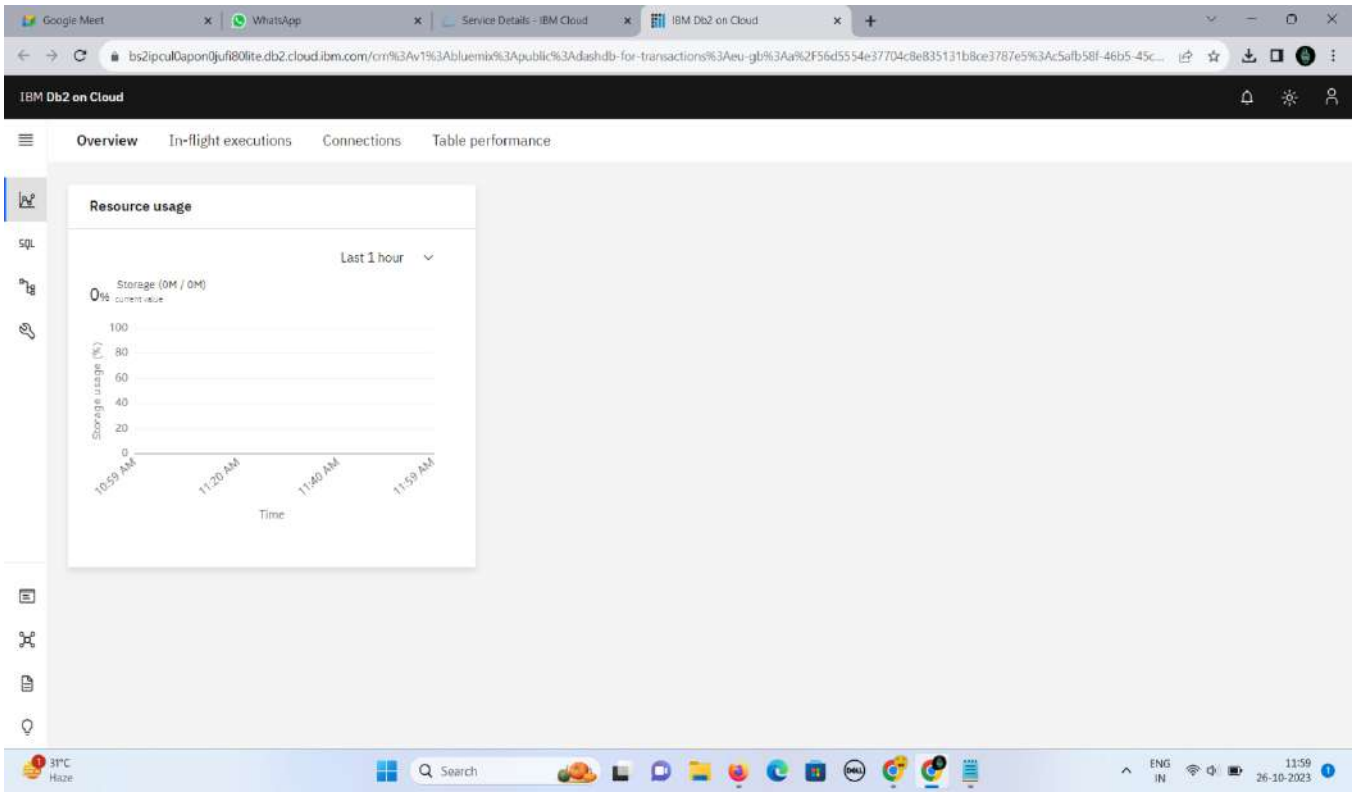


Step 6:

*click “Go To UI” for your data file upload

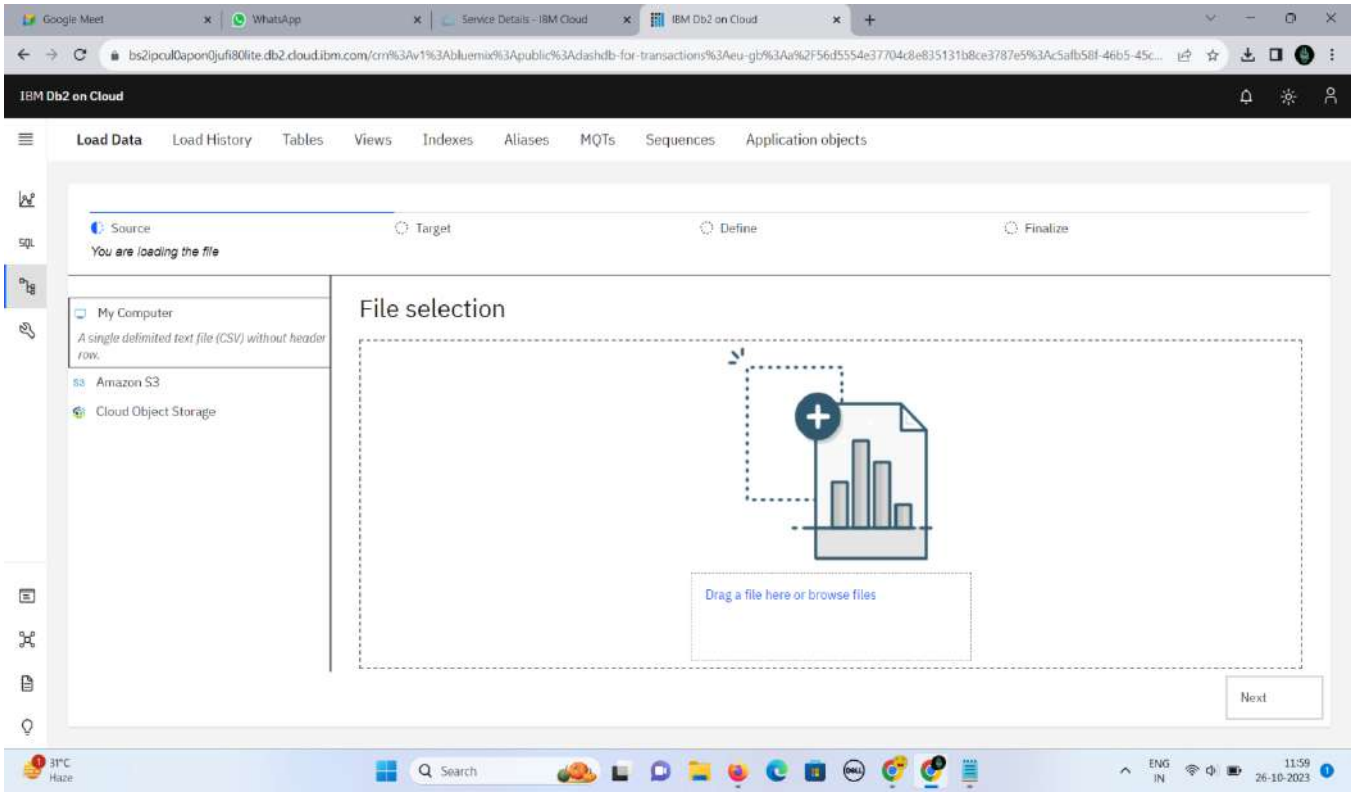


*select your Load Data

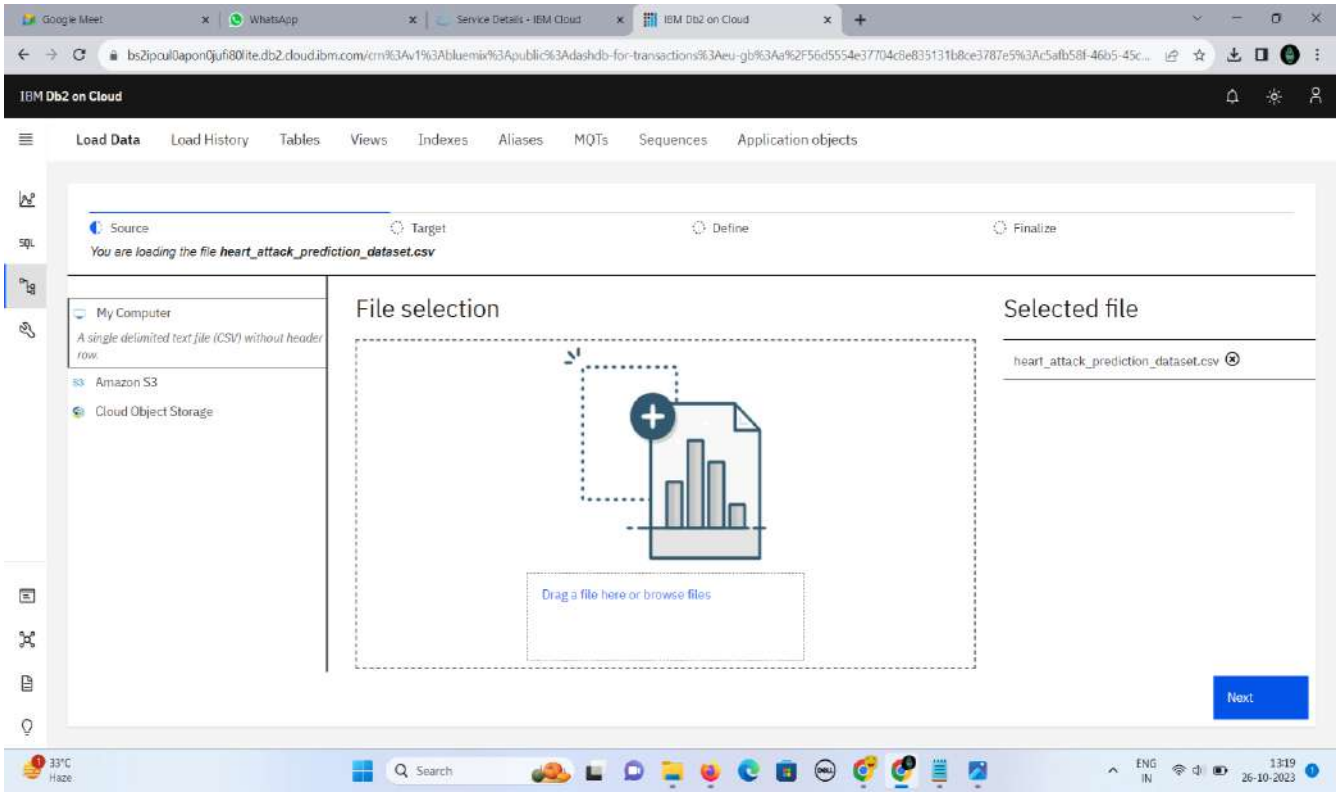


Step 7:

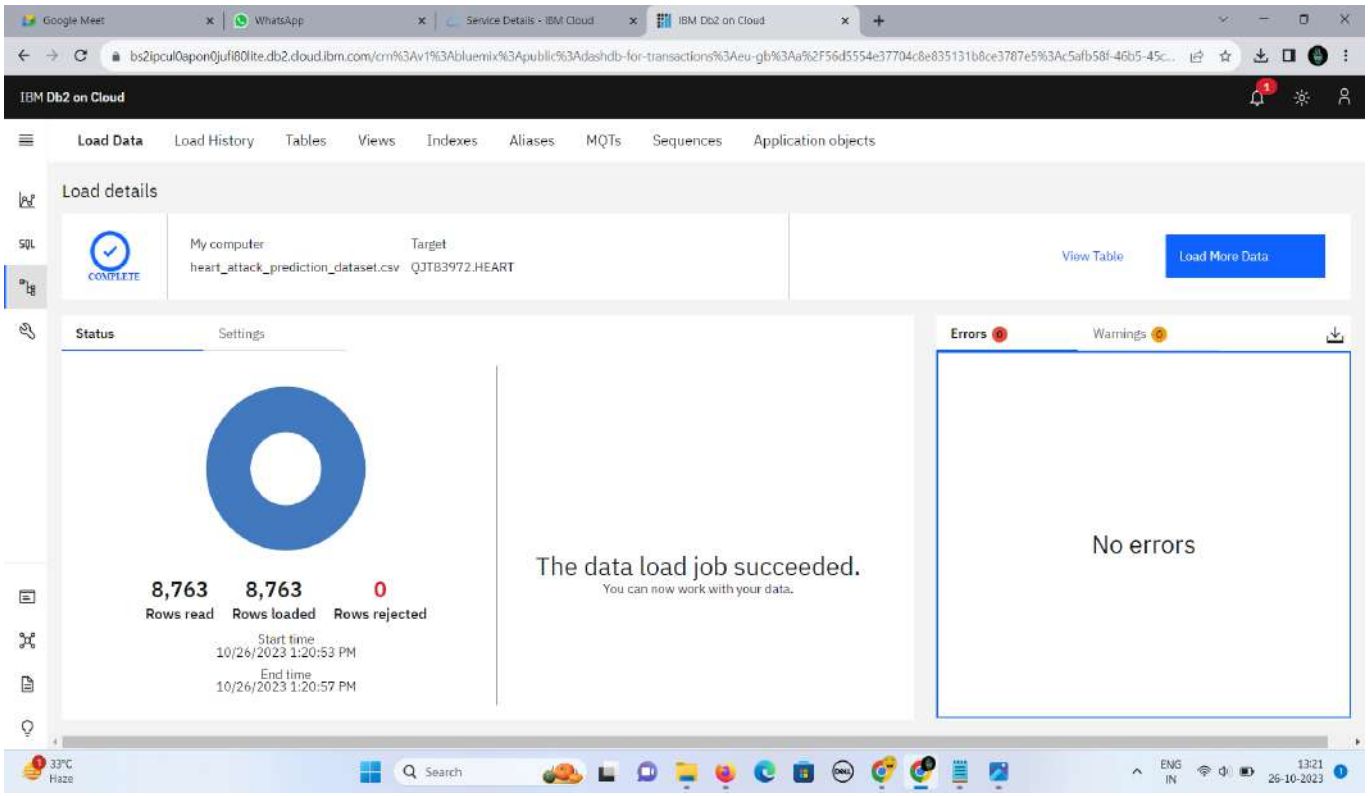
*uplaod our csv file data from “My Computer”



*upload our csv file (“heart_attack_prediction_dataset.csv”)



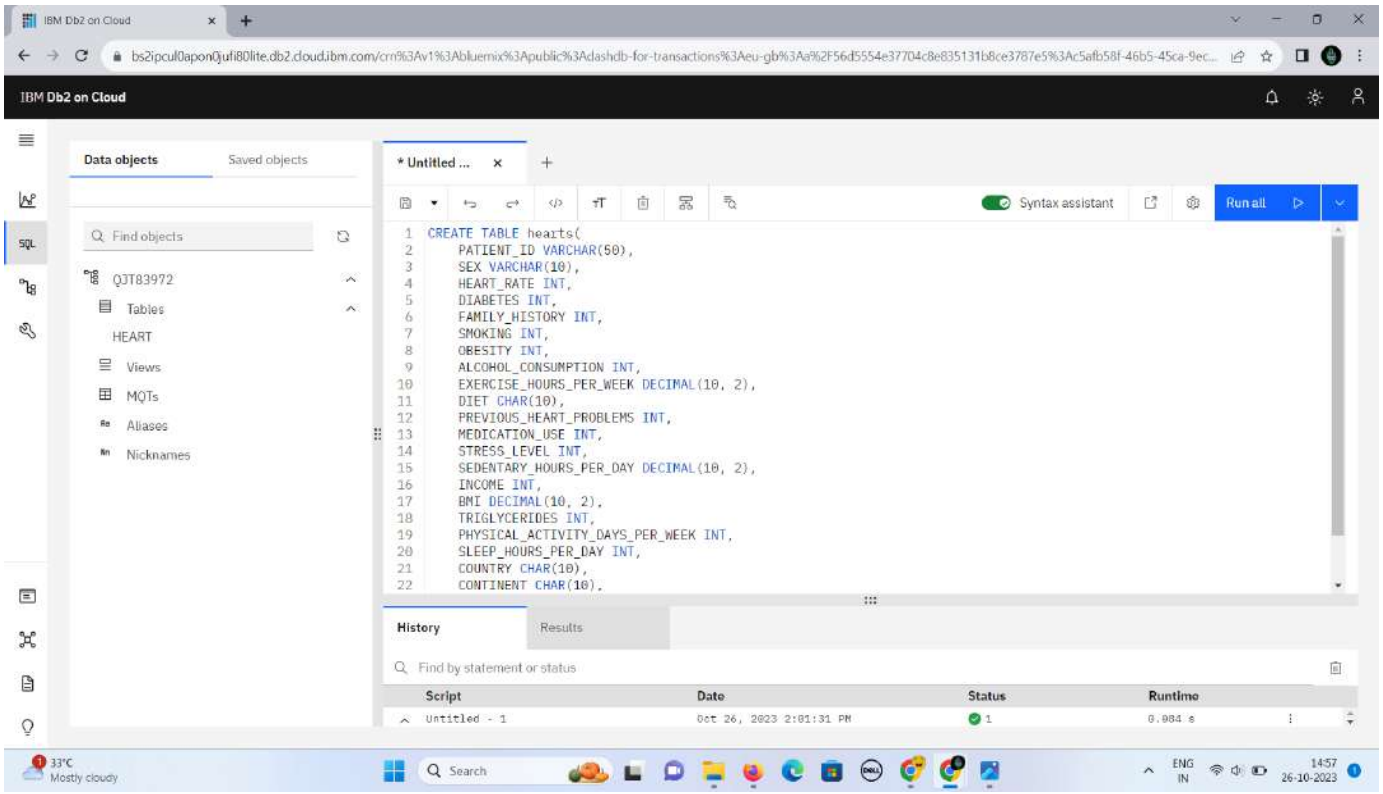
*The server read your data completely and upload your data on your server and completing your data upload process



Step 8:

*After completely uploading our dataset

*Go To SQL section and execute Data Cleaning and transformation queries



3.2. Development Part 2

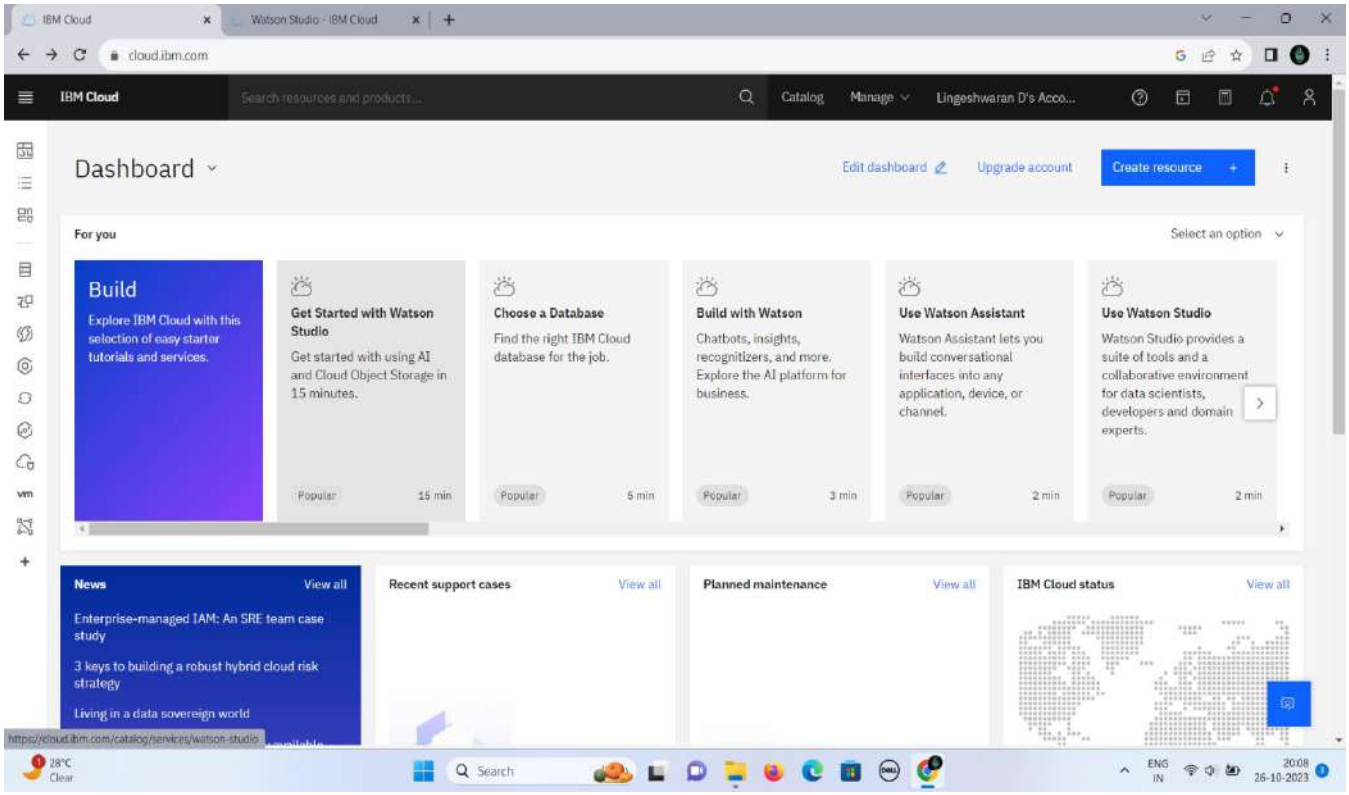
Building on the foundations laid in the first phase, we will deepen our analysis. This phase includes applying advanced analysis techniques and creating more comprehensive visualizations. It's where the true power of big data analysis comes to the forefront.

This part contains several steps.

Step 1:

*login our IBM Cloud Account and create ibm watson studio

*search IBM Watson studio on search bar

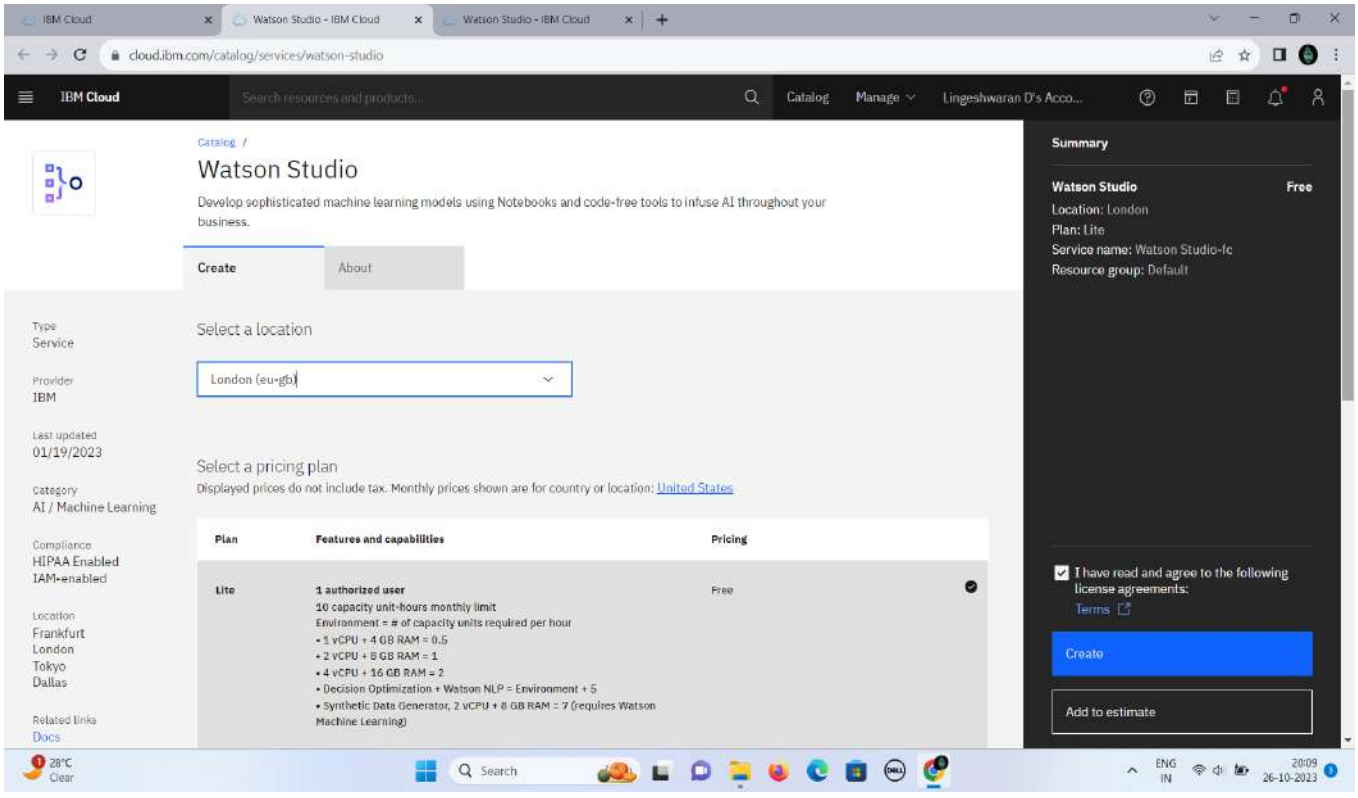


Step 2:

*Create a watson studio and select a location

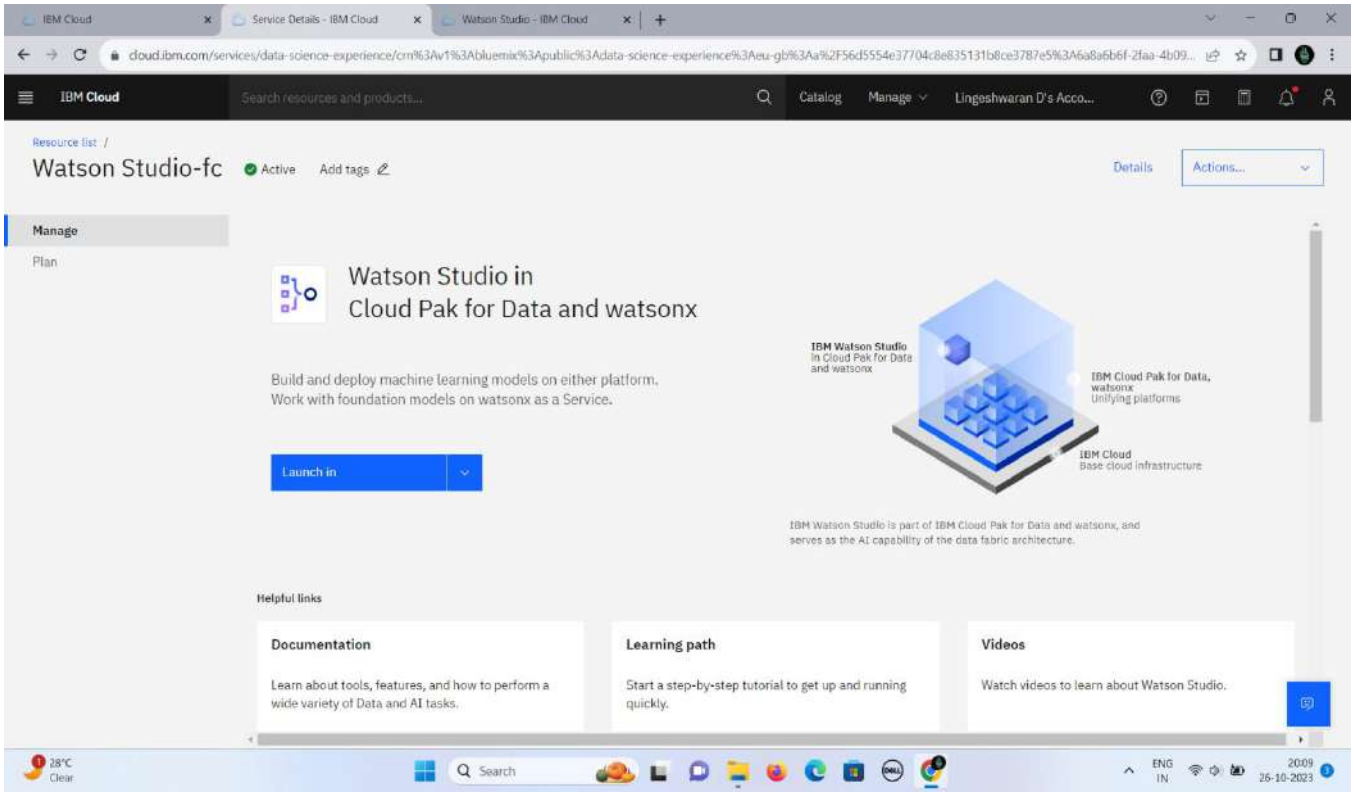
*After selecting the location and click create

Figure 1:

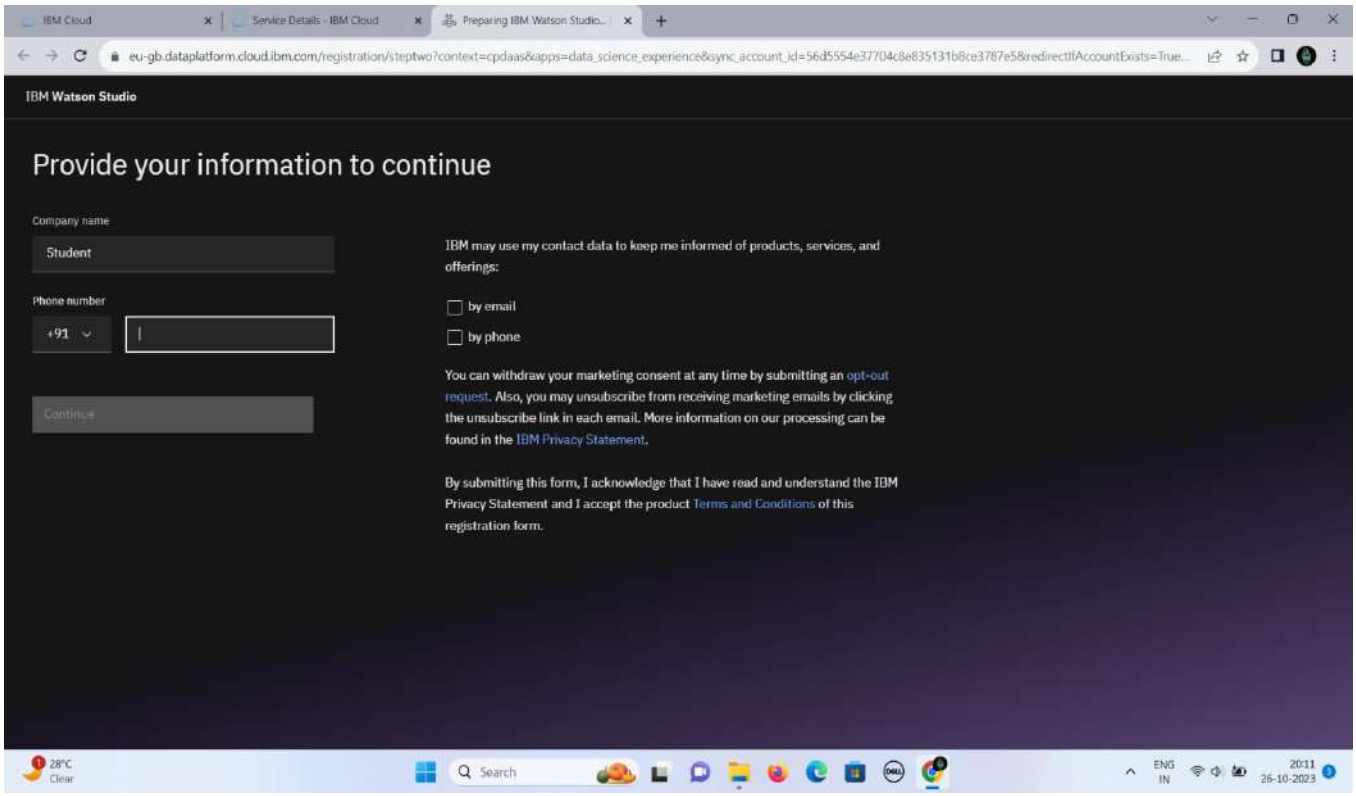


*launch the IBM Watson Studio

Figure 2:

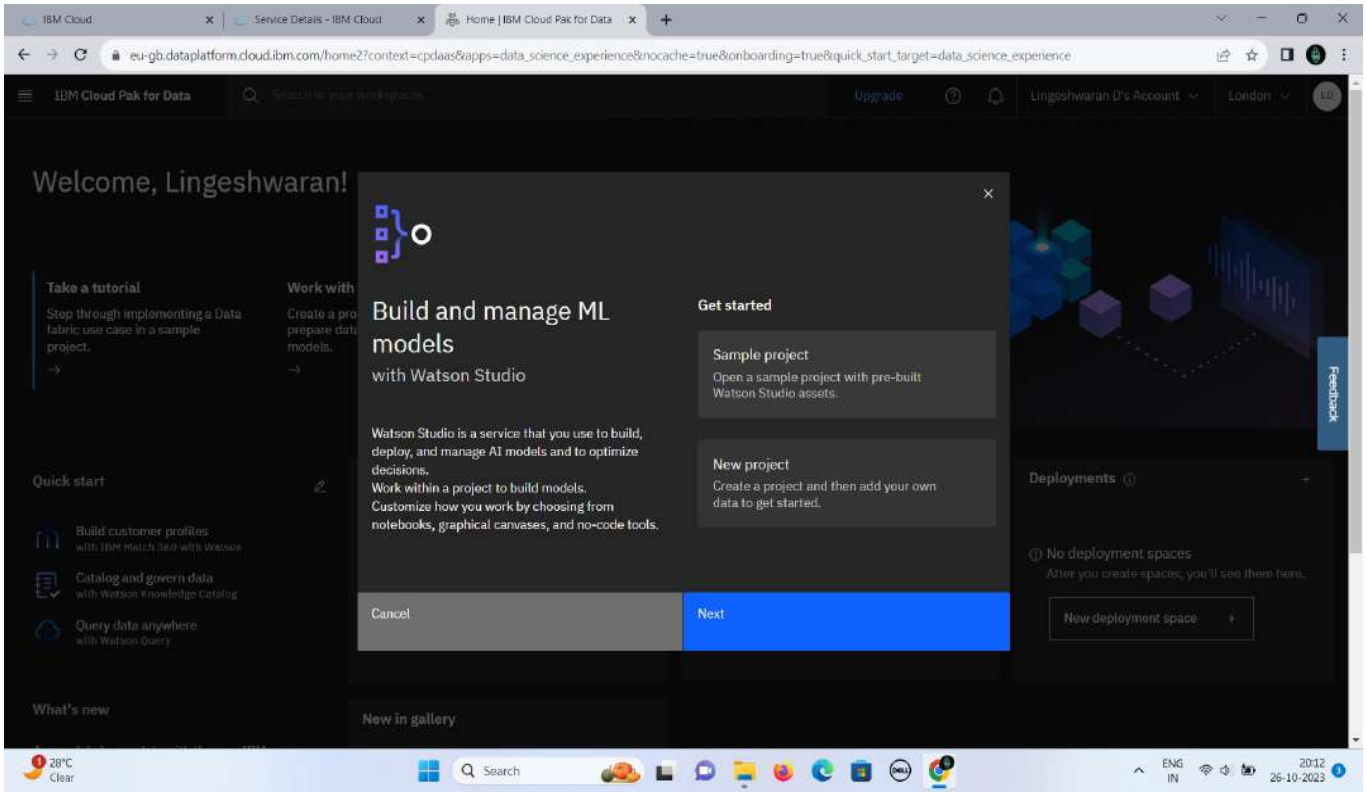


*Enter your information to continue

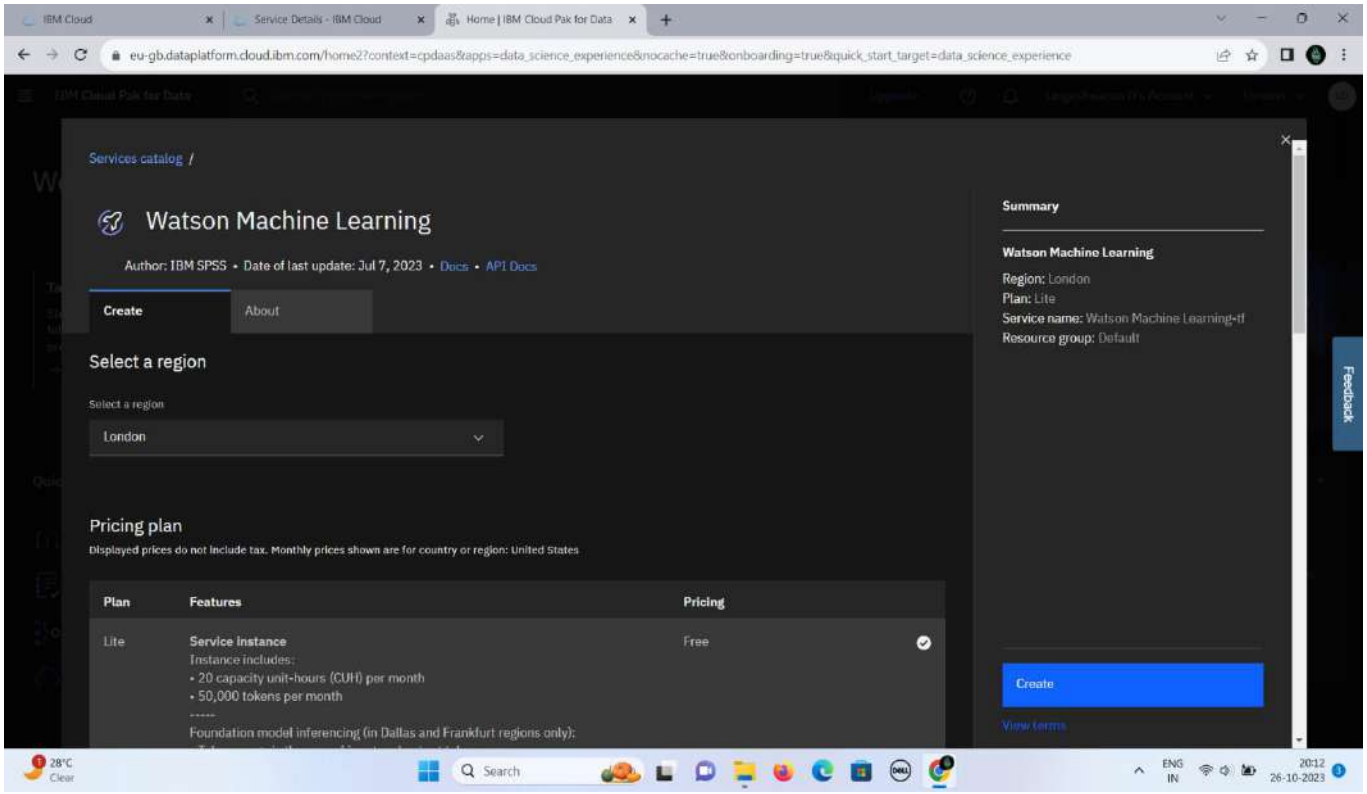


Build and manage ML model with watson studio

Watson studio is a service that you use to build,deplay and manage AI model and to optimise decision work within a project to build model customize how you work by choosing from notebook,graphical canvases and no code tools

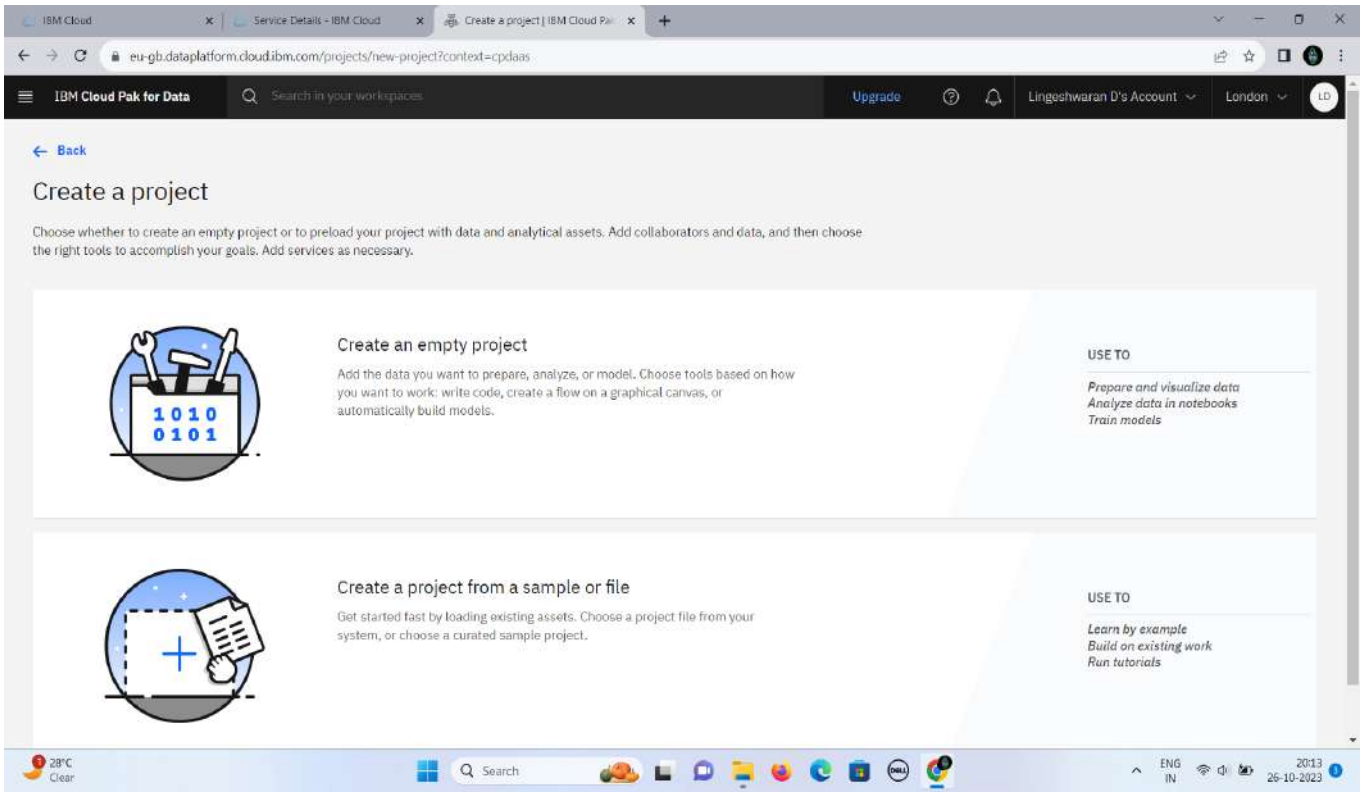


*create watson machine learning in IBM watson studio



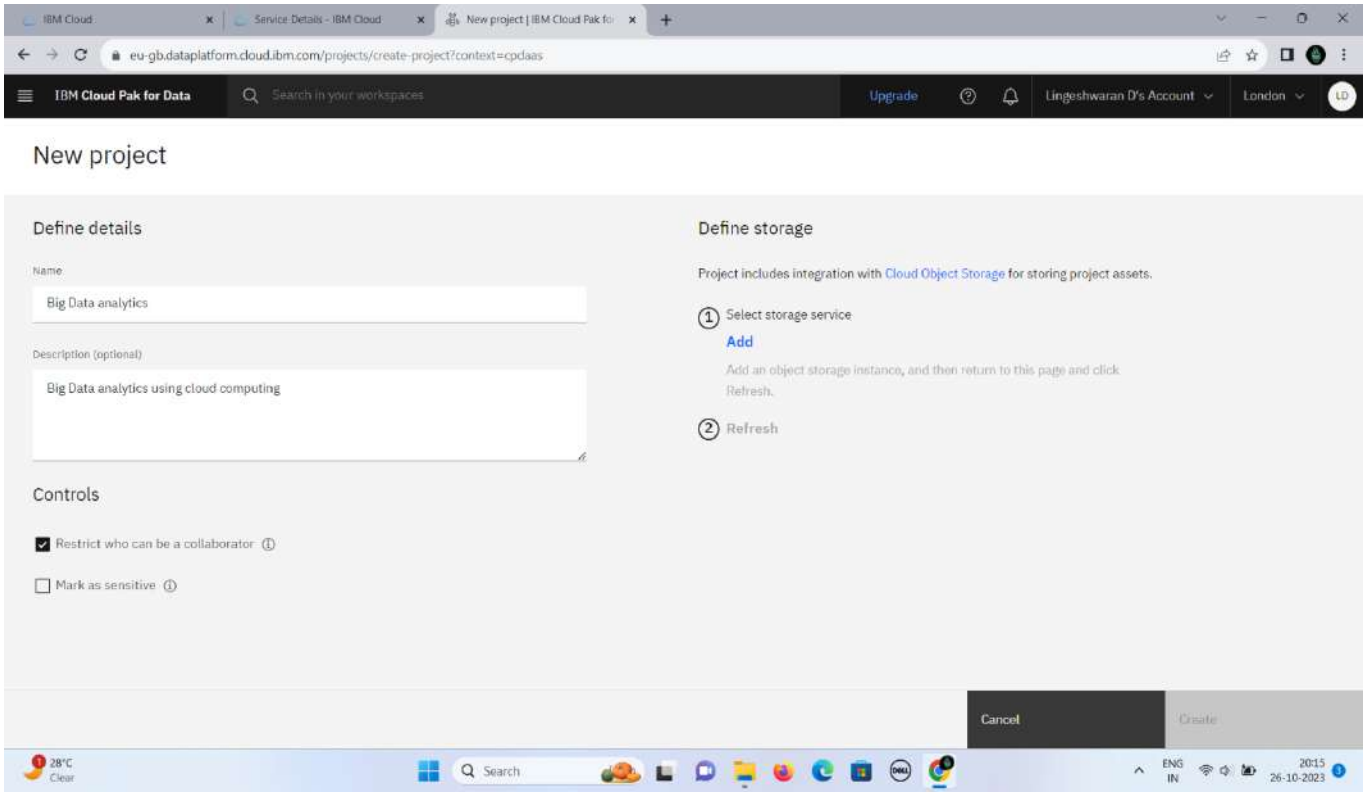
***create a project to work on IBM watson studio**

Choose whether to create an empty project or to preload with date and analytical assets.add collaborators and date,and then choose the right tools to accomplish your goals.add service as necessary

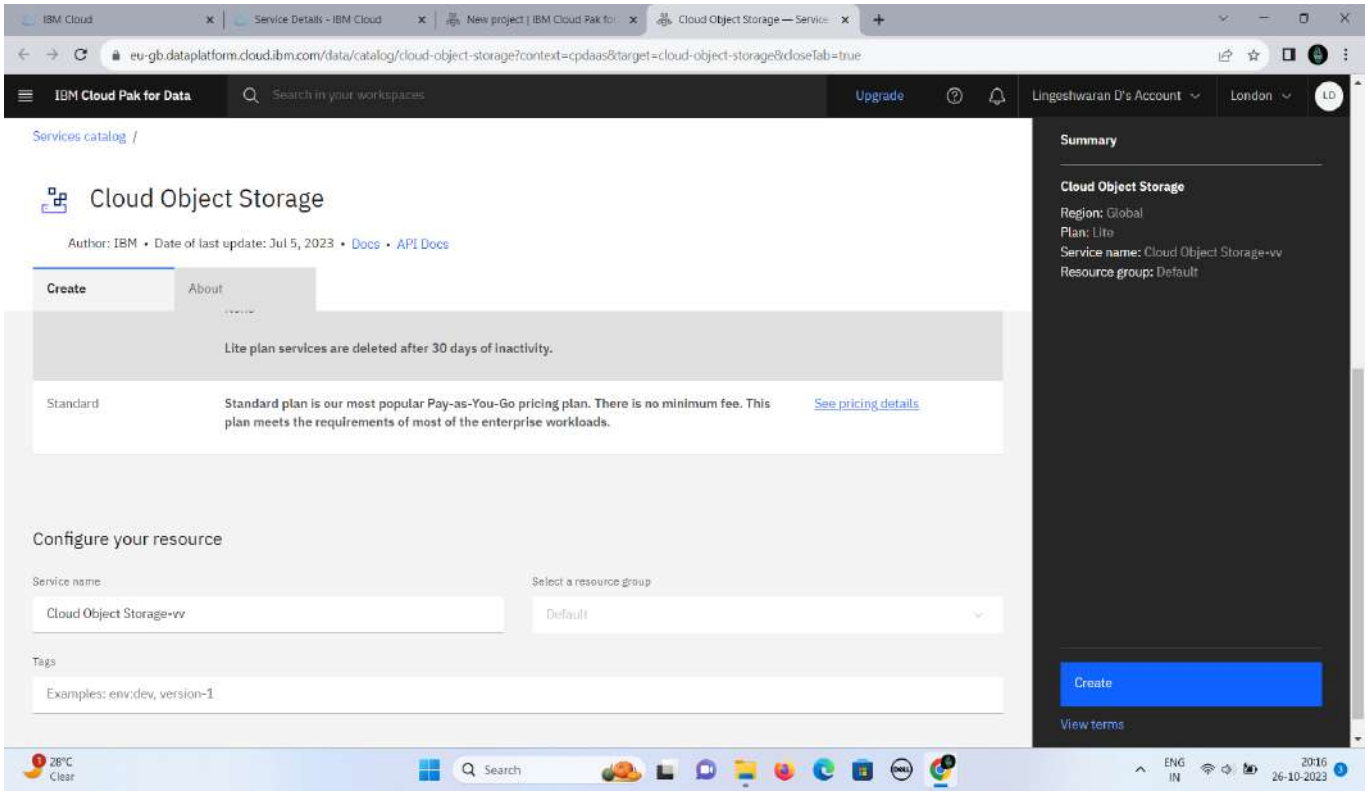


*Give your project name and description of your project

*Add free storage on IBM watson studio for our dataset (“Cloud Object Storage”)



*”cloud object storage” purchase process for storing a dataset



* After complete a name ,description & cloud object storage process and create your project

Figure 1:

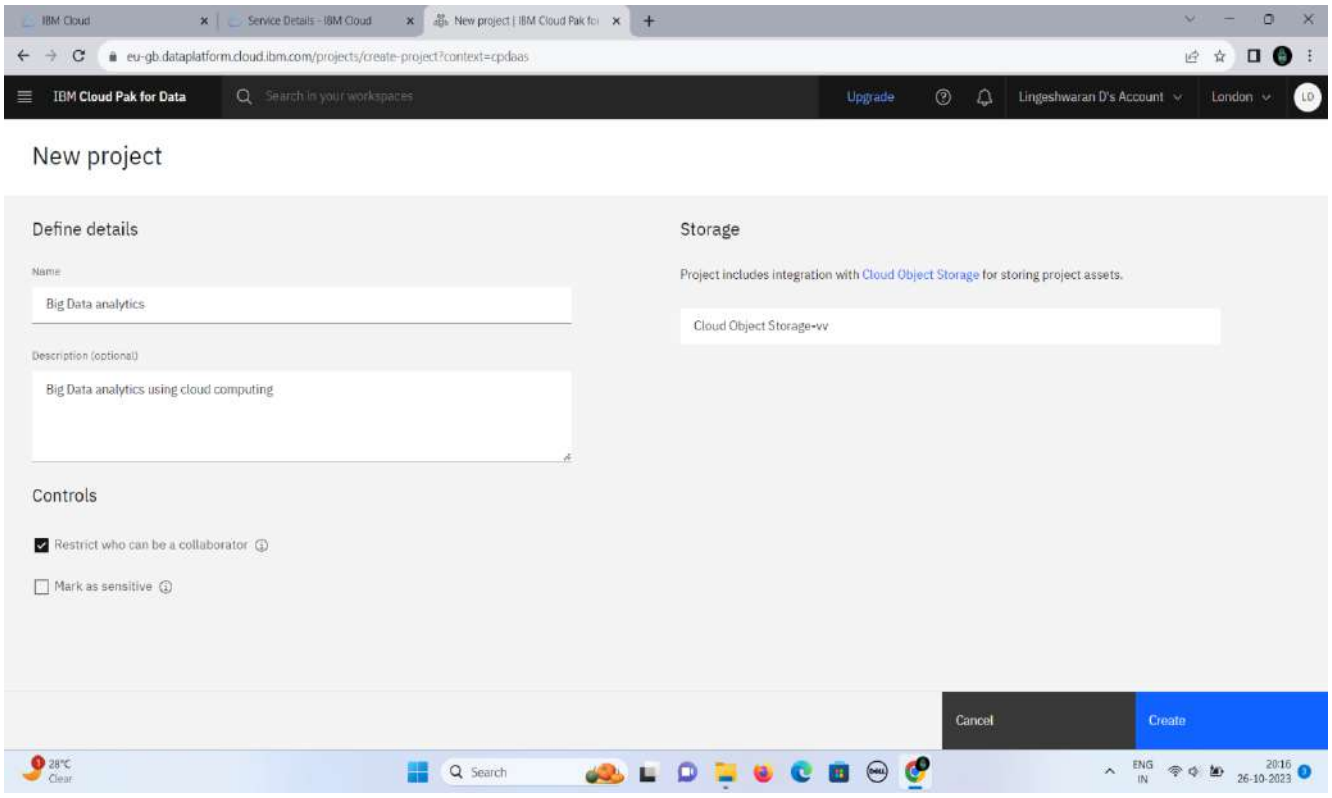
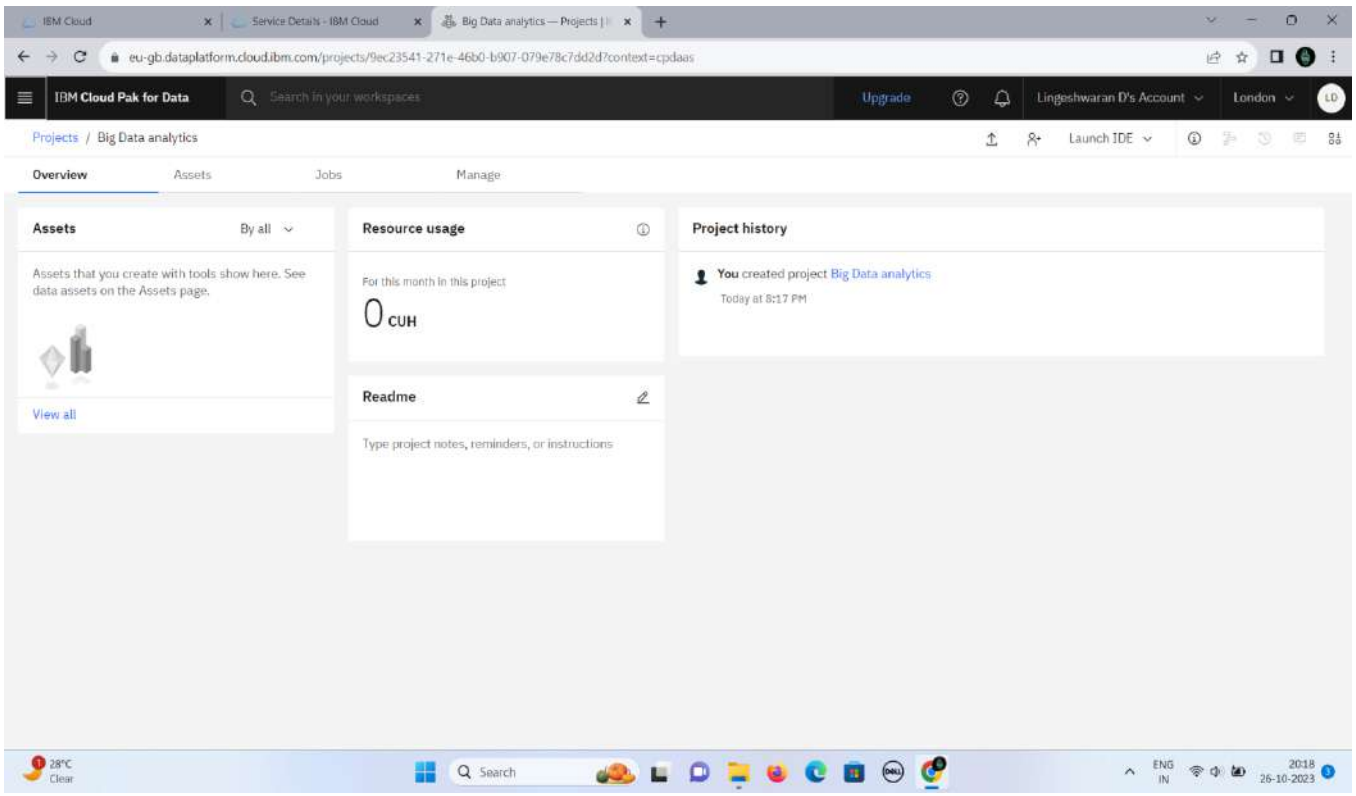


Figure 2:

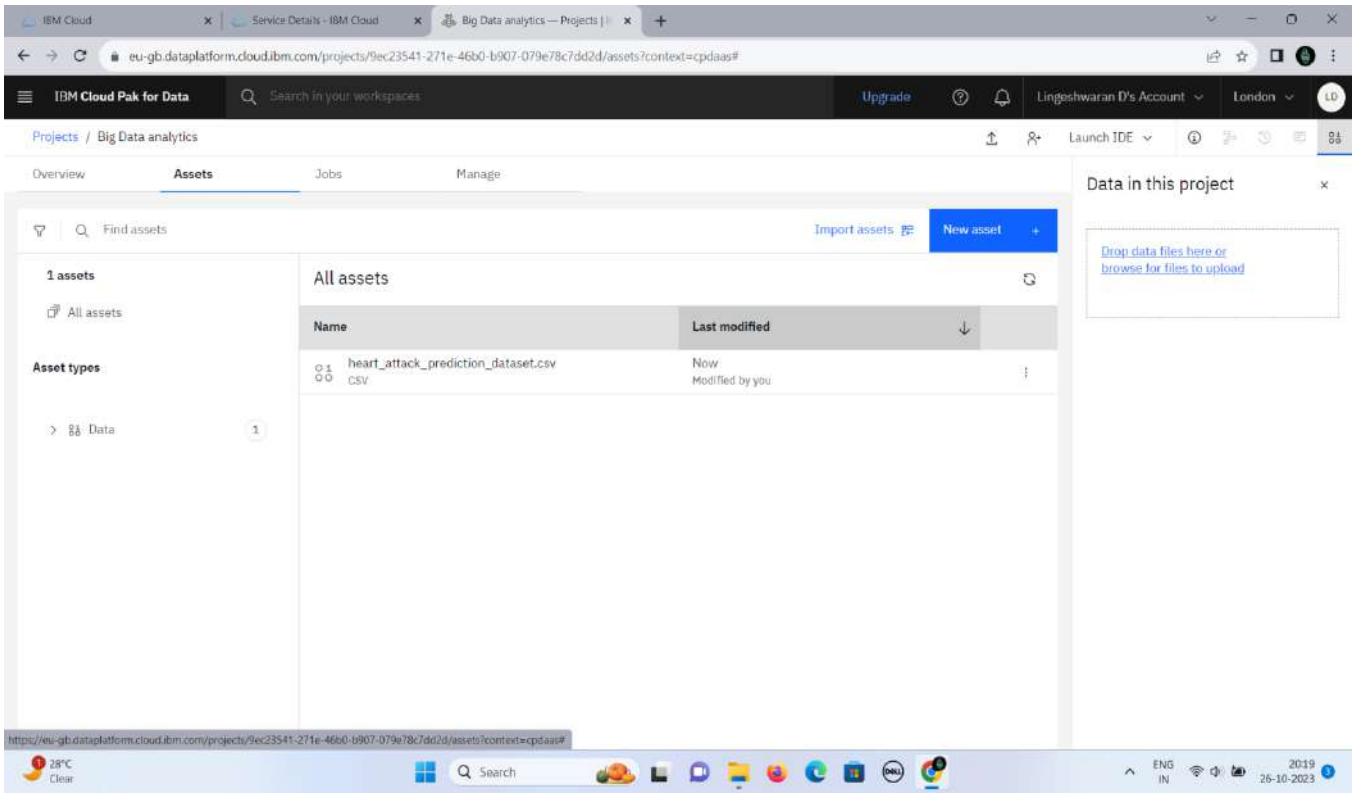

IBM Cloud
Service Details - IBM Cloud
Big Data analytics — Projects
eu-gb.dataplatform.cloud.ibm.com/projects/9ec23541-271e-46b0-b907-079e78c7dd2d?context=cpdaas
IBM Cloud Pak for Data
Search in your workspaces
Upgrade
Lingeshwaran D's Account
London
Projects / Big Data analytics
Launch IDE
Overview
Assets
Jobs
Manage
Assets
By all
Assets that you create with tools show here. See data assets on the Assets page.
View all
Resource usage
For this month in this project
0 CUH
Readme
Type project notes, reminders, or instructions
Project history
You created project Big Data analytics
Today at 8:17 PM
28°C
Clear
Search
ENG
IN
20:18
26-10-2023

*Upload your csv dataset file on IBM watson studio (“heart_attack_prediction_dataset.csv”)

*click Assets to view our csv file

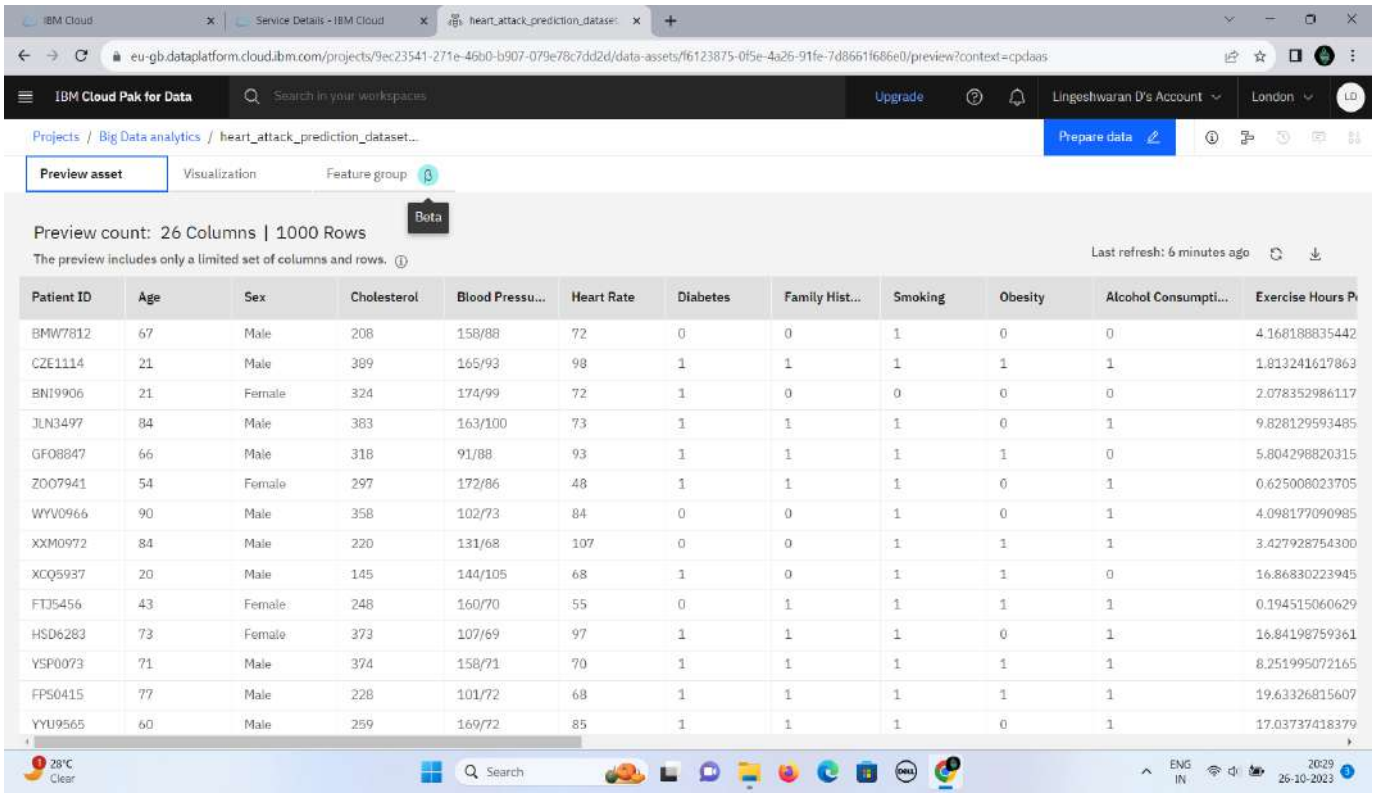
*click on your csv file for preview asset

Figure 3:



*select a visualizaiton on your csv.file

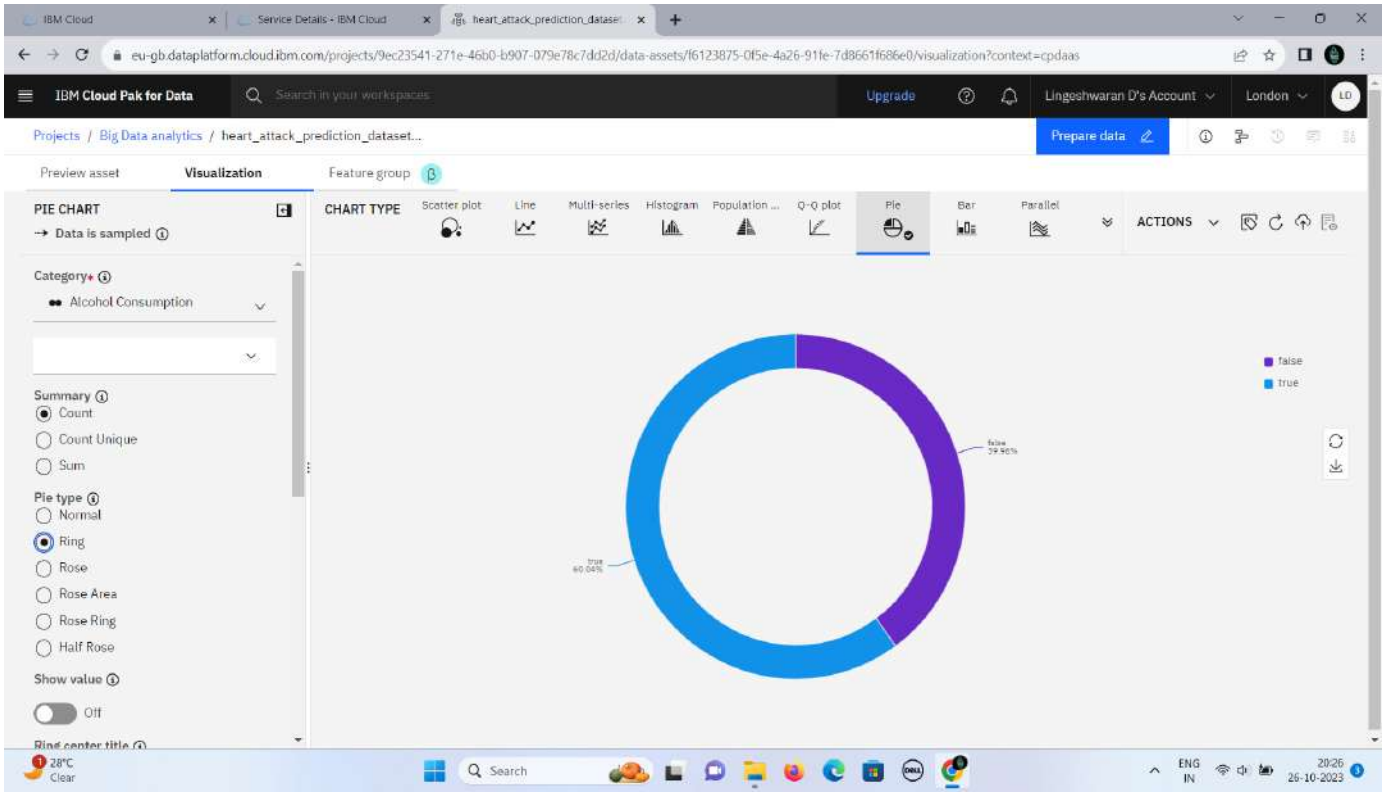
Figure 4:



Step 3:

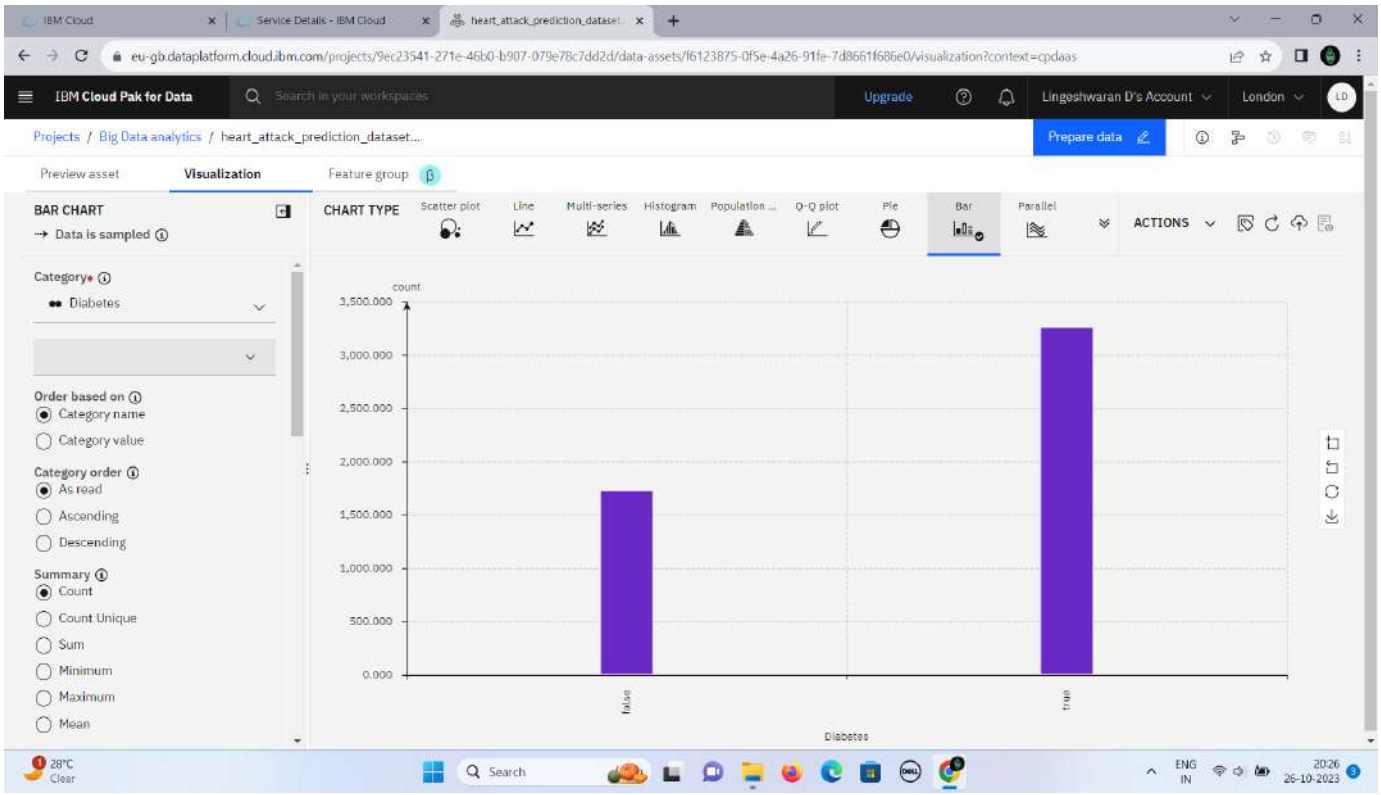
Visualization our dataset using IBM Watson studio

PIECHART:



BAR CHART:

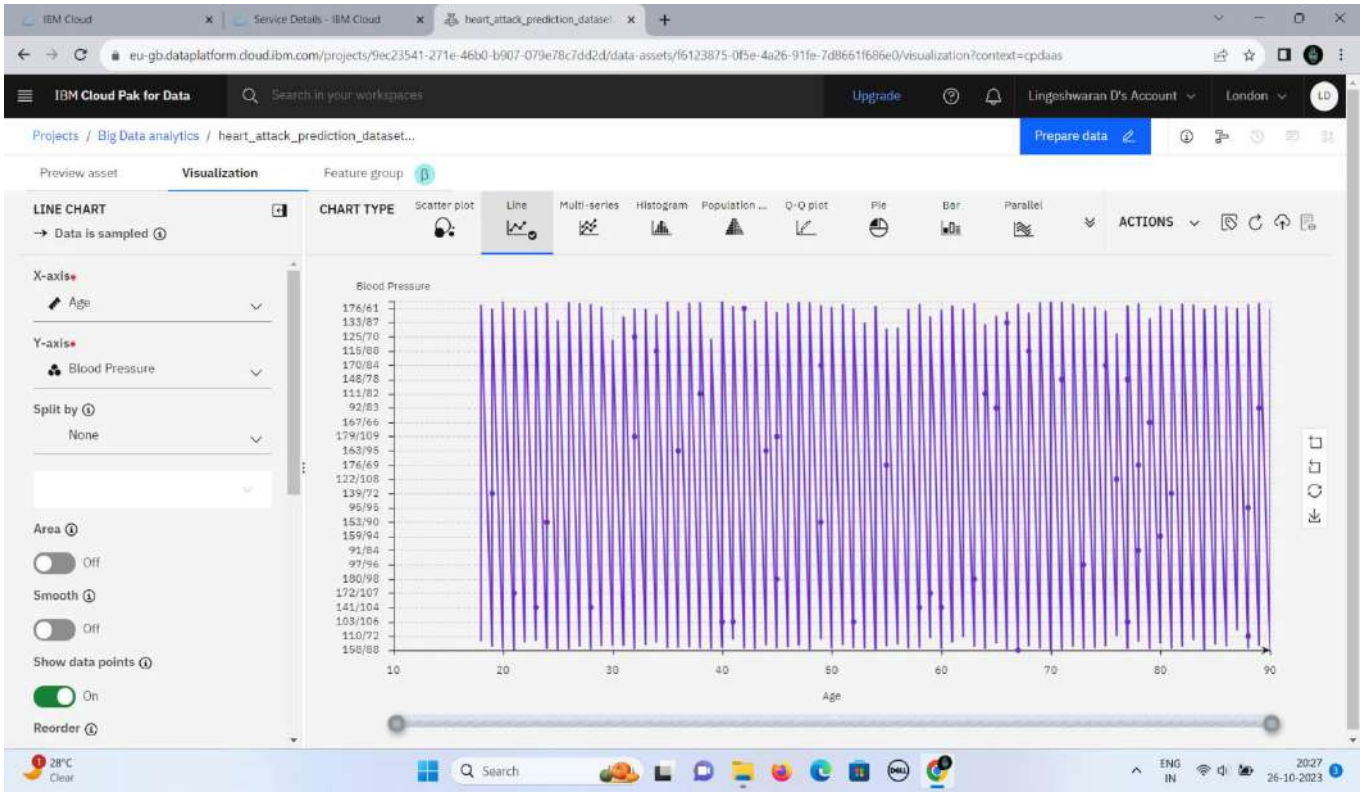
*For Diabetes



LINE CHART:

*X-AXIS(Age)

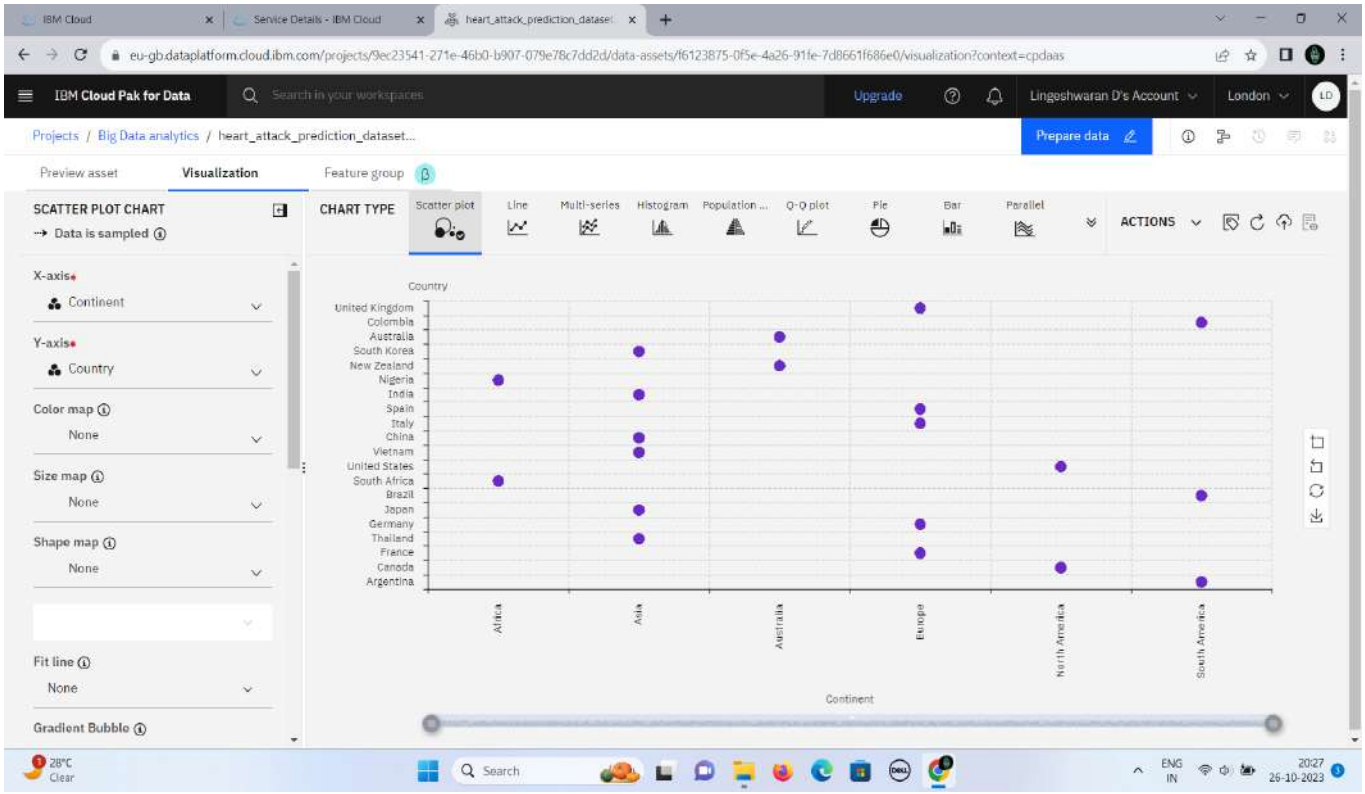
*Y-AXIS(Blood pressure)



SCATTER PLOT CHART:

*X-AXIS(CONTINENT)

*Y-AXIS(COUNTRY)

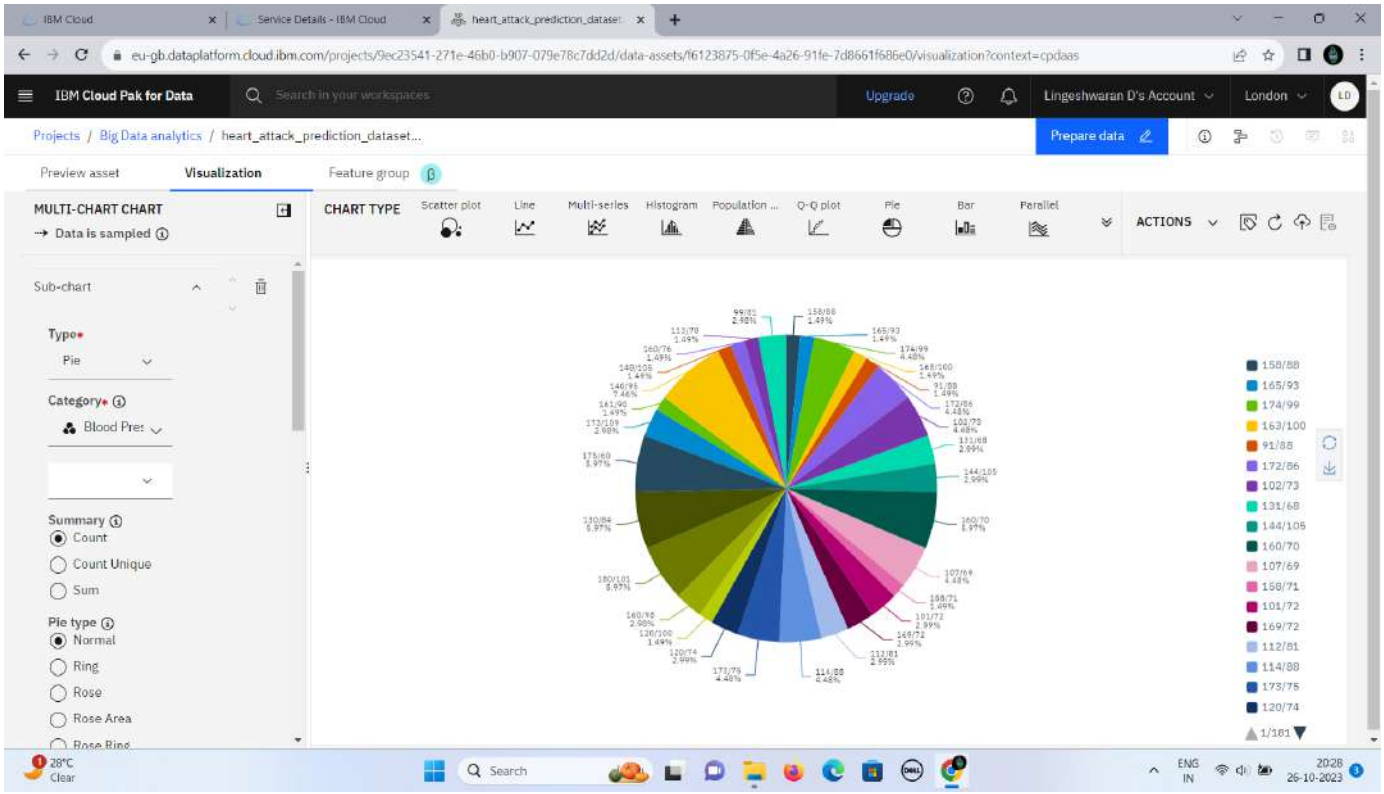


MULTI_CHART CHART:

SUB-CHART

*CATEGORY=BLOOD PRESSURE

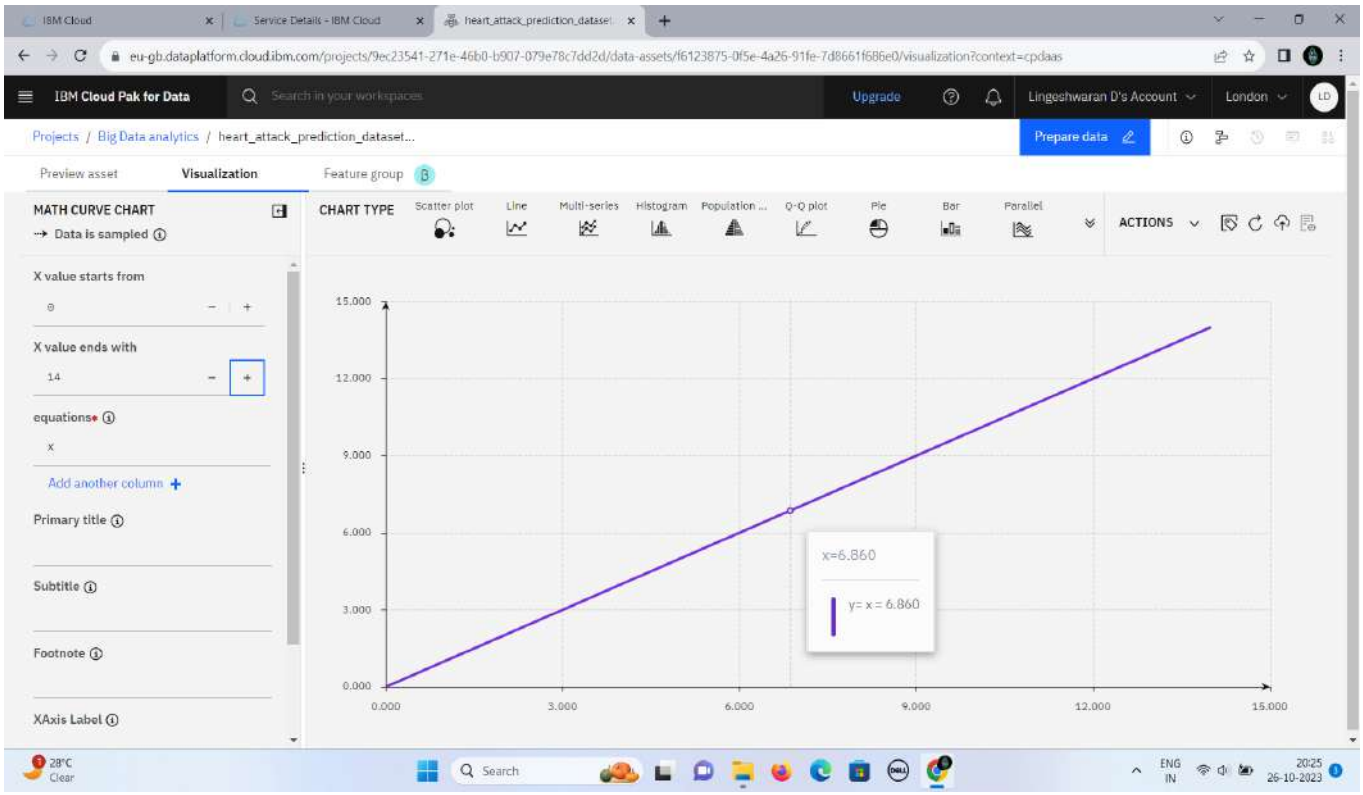
*TYPE=PIE



MATH CURVE CHART:

*X-AXIS(0)

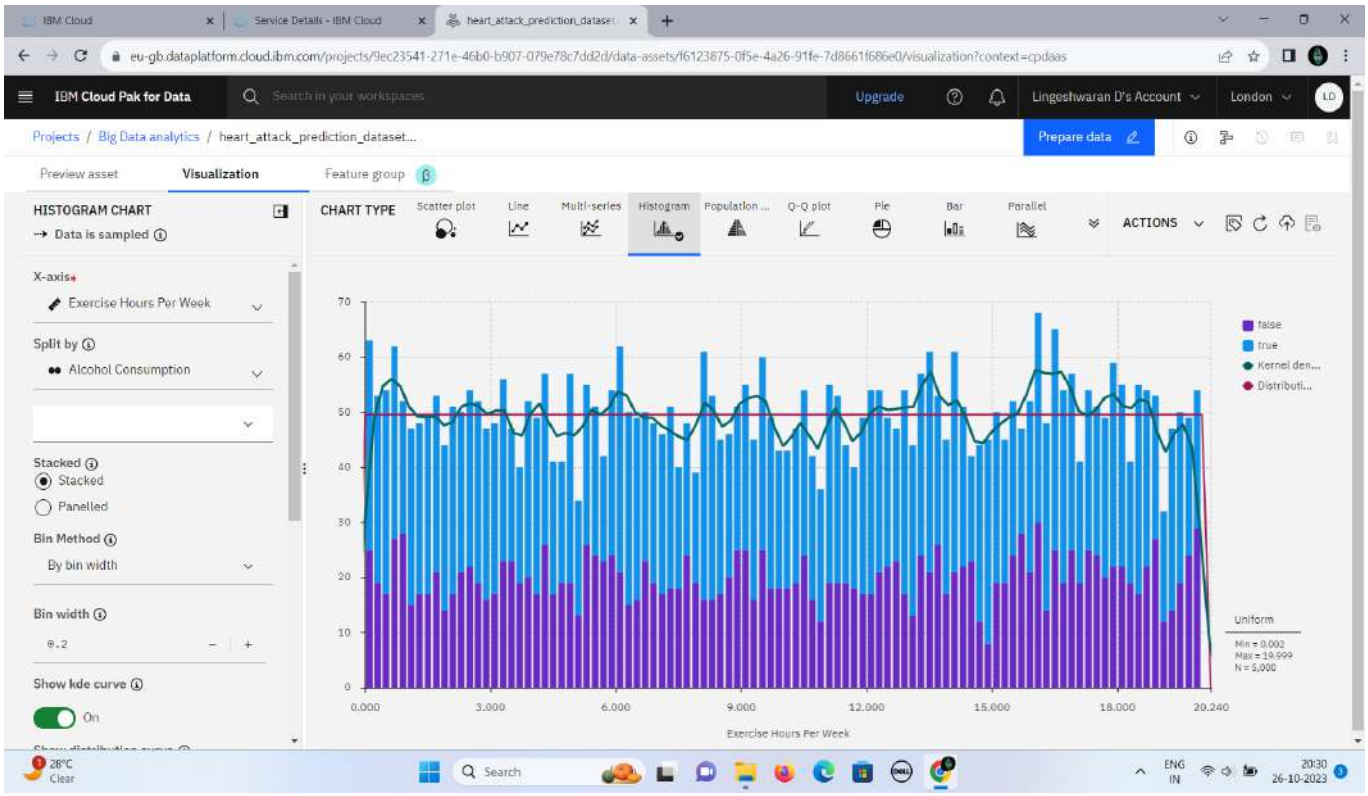
*Y-AXIS(14)



HISTOGRAM CHART:

*X-AXIS(EXERCISE HOURS PER WEEK)

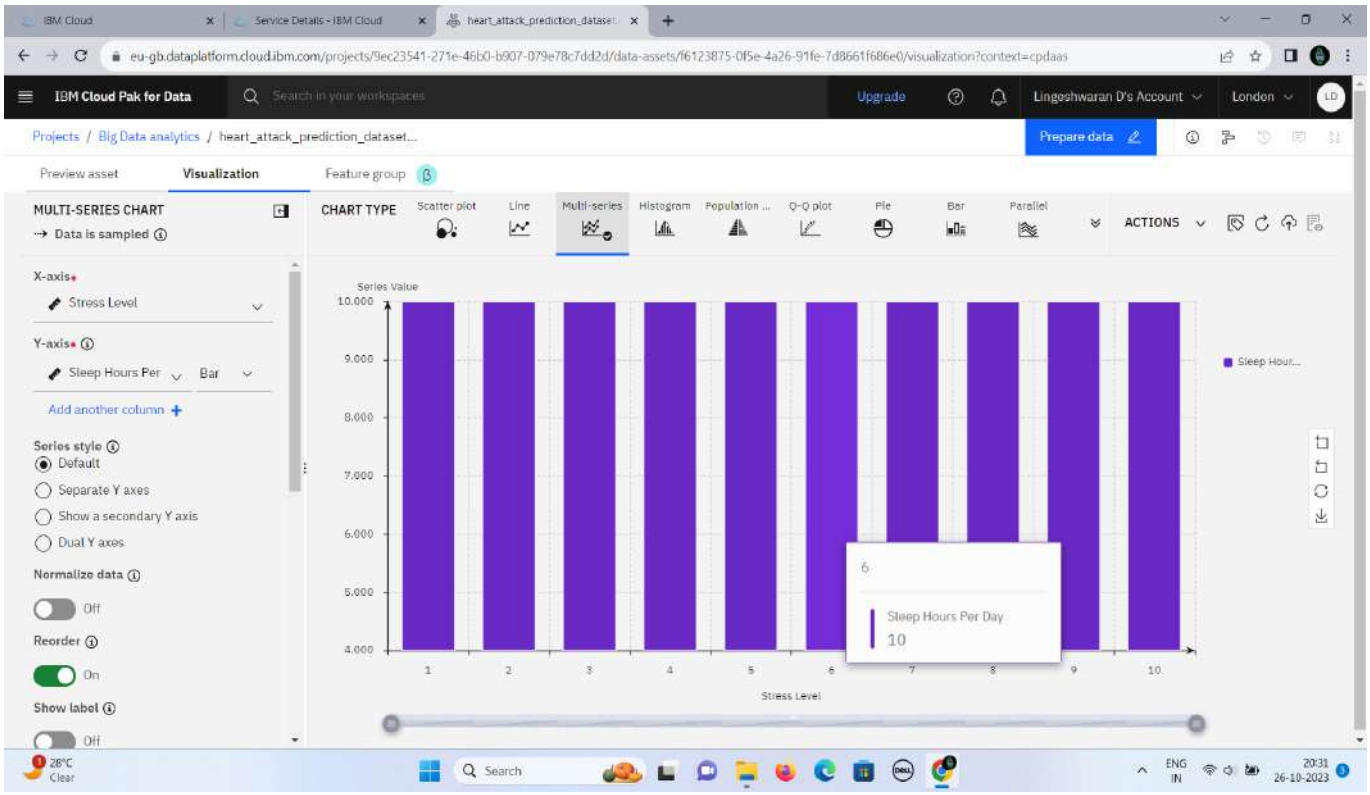
*SPLIT BY(ALCOHOL CONSUMPTION)



MULTI-SERIES CHART:

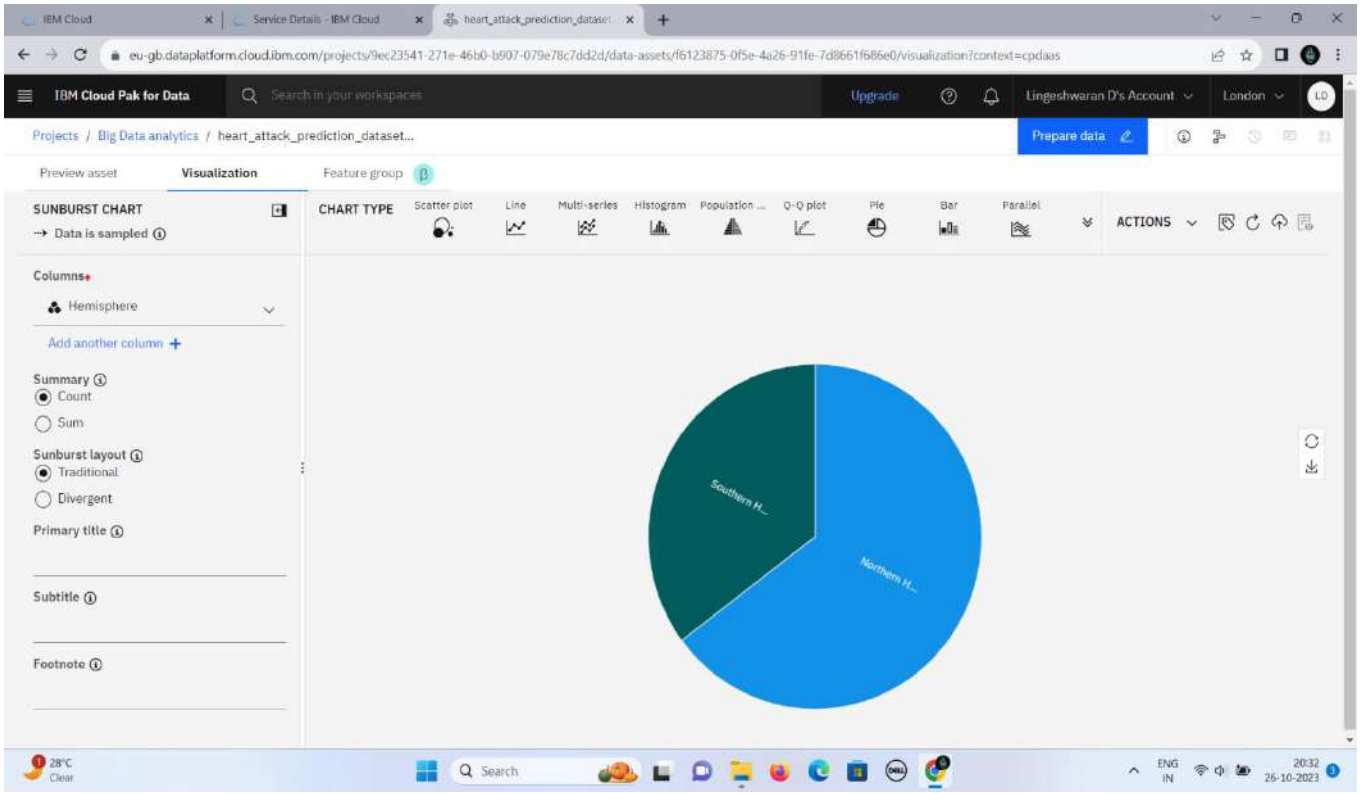
*X-AXIS(STRESS LEVEL)

*Y-AXIS(SLEEP HOURS PER WEEK)



SUNBURST CHART :

COLUMNS=HEMISPHERE



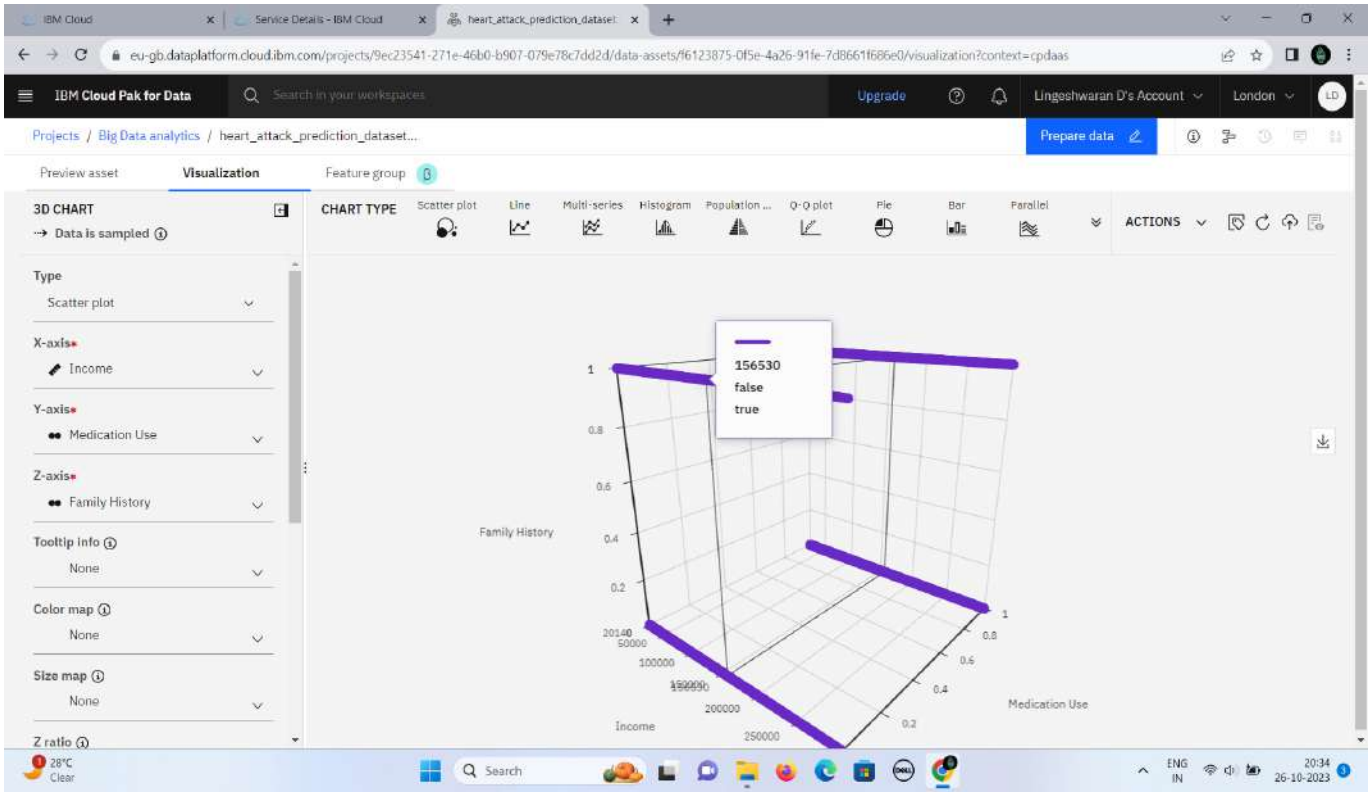
3D CHART:

*TYPE()

*X-AXIS(INCOME)

*Y-AXIS(MEDICINE USE)

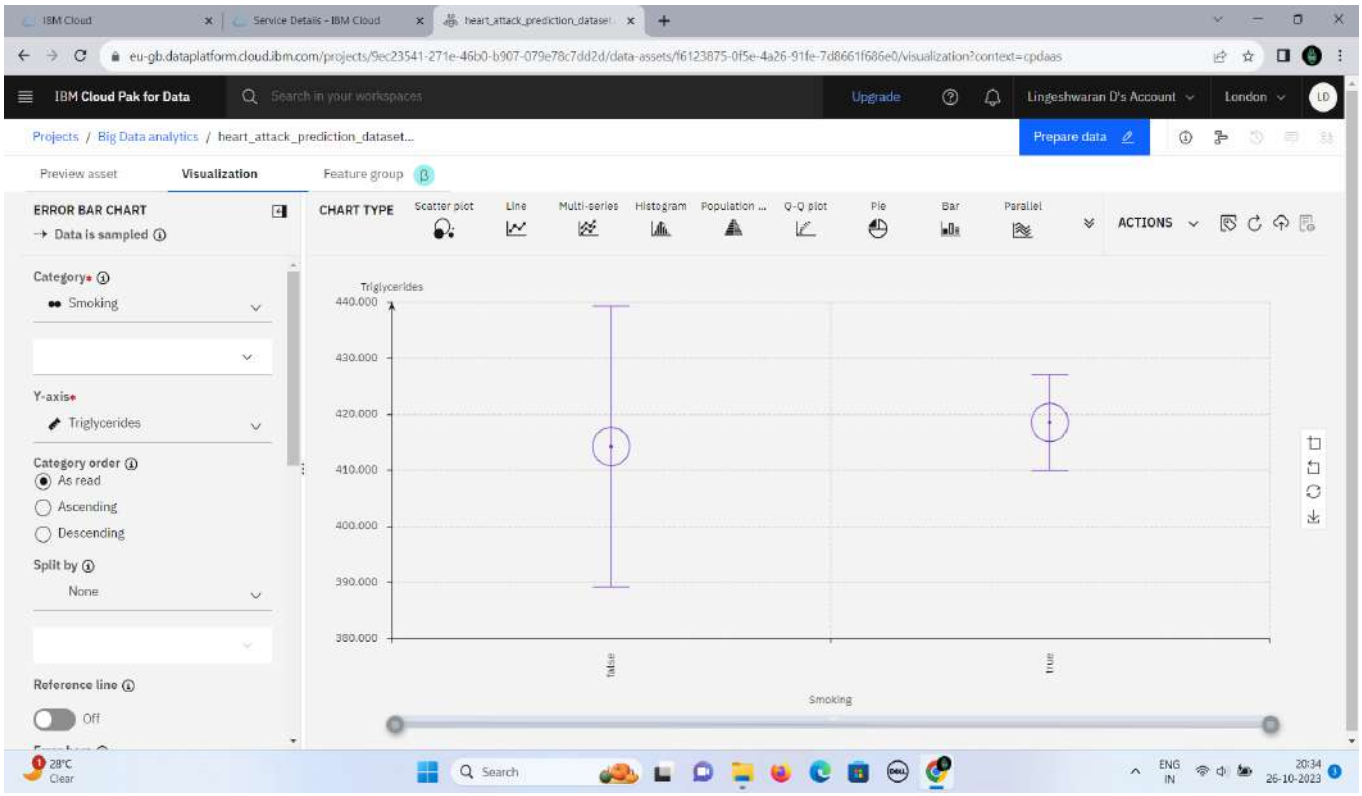
*Z-AXIS(FAMILY HISTORY)



ERROR BAR CHART:

*CATEGORY(SMOKING)

*Y-AXIS(TRIGLYCERIDES)



Big data analytics solution by applying advanced analysis techniques using python code

IBM cloud notebook:

```
!pip install pandas scikit-learn
Requirement already satisfied: pandas in
/opt/conda/envs/Python-3.10/lib/python3.10/site-packages (1.4.3)
Requirement already satisfied: scikit-learn in /opt/conda/envs/Python3.10 /lib/python3.10/site-packages (1.1.1)
Requirement already satisfied: python-dateutil>=2.8.1 in
/opt/conda/envs/Python-3.10/lib/python3.10/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/envs/Python3.10 /lib/python3.10/site-packages (from
pandas) (2023.3.post )
Requirement already satisfied: numpy>=1.21.0 in
/opt/conda/envs/Python-3.10/lib/python3.10/site-packages (from pandas)(1.23.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/envs/Python-3.10/lib/python3.10/site-packages (from scikitlearn) (2.2.0)
Requirement already satisfied: joblib>=1.0.0 in
/opt/conda/envs/Python-3.10/lib/python3.10/site-packages (from scikitlearn) (1.1.1)
Requirement already satisfied: scipy>=1.3.2 in /opt/conda/envs/Python-
3.10 /lib/python3.10/site-packages (from scikit-learn ) (1.8.1)
Requirement already satisfied: six>=1.5 in /opt/conda/envs/Python3.10/lib/python3.10/site-packages (from python-
dateutil>=2.8.1> pandas ) (1.16.0)

#loading the dataset

import os, types import pandas as pd from
botocore.client import Config import ibm_boto3 def
__iter__(self): return 0

# @hidden_cell

# The following code accesses a file in your IBM Cloud Object Storage.

It includes your credentials.

# You might want to remove those credentials before you share the notebook. cos_client =
ibm_boto3.client(service_name='s3',

   ibm_api_key_id='bNtKn-suJyy3PdlnIbNENRb7ZoRxjw0Bk6tjKdfwXk9H',
ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",    config=Config(signature_version='oauth') ,
endpoint_url='https://s3.private.eu-gb.cloud-objectstorage.appdomain.cloud')

bucket = 'bigdataanalyticsusingcloudcomputi-donotdelete-prbnssfschsf5bmp'

object_key = 'heart_attack_prediction_dataset.csv'

body = cos_client.get_object(Bucket=bucket,Key=object_key)['Body'] # add missing __iter__ method, so pandas
accepts body as file-like object
```

```
df = pd.read_csv(body)
df.head()
```

```
  Patient ID  Age     Sex  Cholesterol Blood Pressure  Heart Rate  Diabetes  \
0    BMW7812   67    Male          208         158/88          72         0
1    CZE1114   21    Male          389         165/93          98         1
2    BNI9906   21  Female          324         174/99          72         1
3    JLN3497   84    Male          383        163/100          73         1
4    GFO8847   66    Male          318          91/88          93         1

   Family History  Smoking  Obesity  ...  Sedentary Hours Per Day  Income  \
0               0        1        0  ...                 6.615001  261404
1               1        1        1  ...                 4.963459  285768
2               0        0        0  ...                 9.463426  235282
3               1        1        0  ...                 7.648981  125640
4               1        1        1  ...                 1.514821  160555

         BMI  Triglycerides  Physical Activity Days Per Week  \
0  31.251233            286                                0
1  27.194973            235                                1
2  28.176571            587                                4
3  36.464704            378                                3
4  21.809144            231                                1

   Sleep Hours Per Day    Country      Continent           Hemisphere  \
0                    6  Argentina  South America  Southern Hemisphere
1                    7     Canada  North America  Northern Hemisphere
2                    4     France         Europe  Northern Hemisphere
3                    4     Canada  North America  Northern Hemisphere
4                    5   Thailand           Asia  Northern Hemisphere

   Heart Attack Risk
0                  0
1                  0
2                  0
3                  0
4                  0

[5 rows x 26 columns]
```

```
df.isnull().sum()
```

```
Patient ID                      0
Age                             0
Sex                             0
Cholesterol                     0
Blood Pressure                  0
Heart Rate                      0
```

```
Diabetes                          0
Family History                    0
Smoking                           0
Obesity                           0
Alcohol Consumption               0
Exercise Hours Per Week           0
Diet                              0
Previous Heart Problems           0
Medication Use                    0
Stress Level                      0
Sedentary Hours Per Day           0
Income                            0
BMI                               0
Triglycerides                     0
Physical Activity Days Per Week   0
Sleep Hours Per Day               0
Country                           0
Continent                         0
Hemisphere                        0
Heart Attack Risk                 0
dtype: int64
```

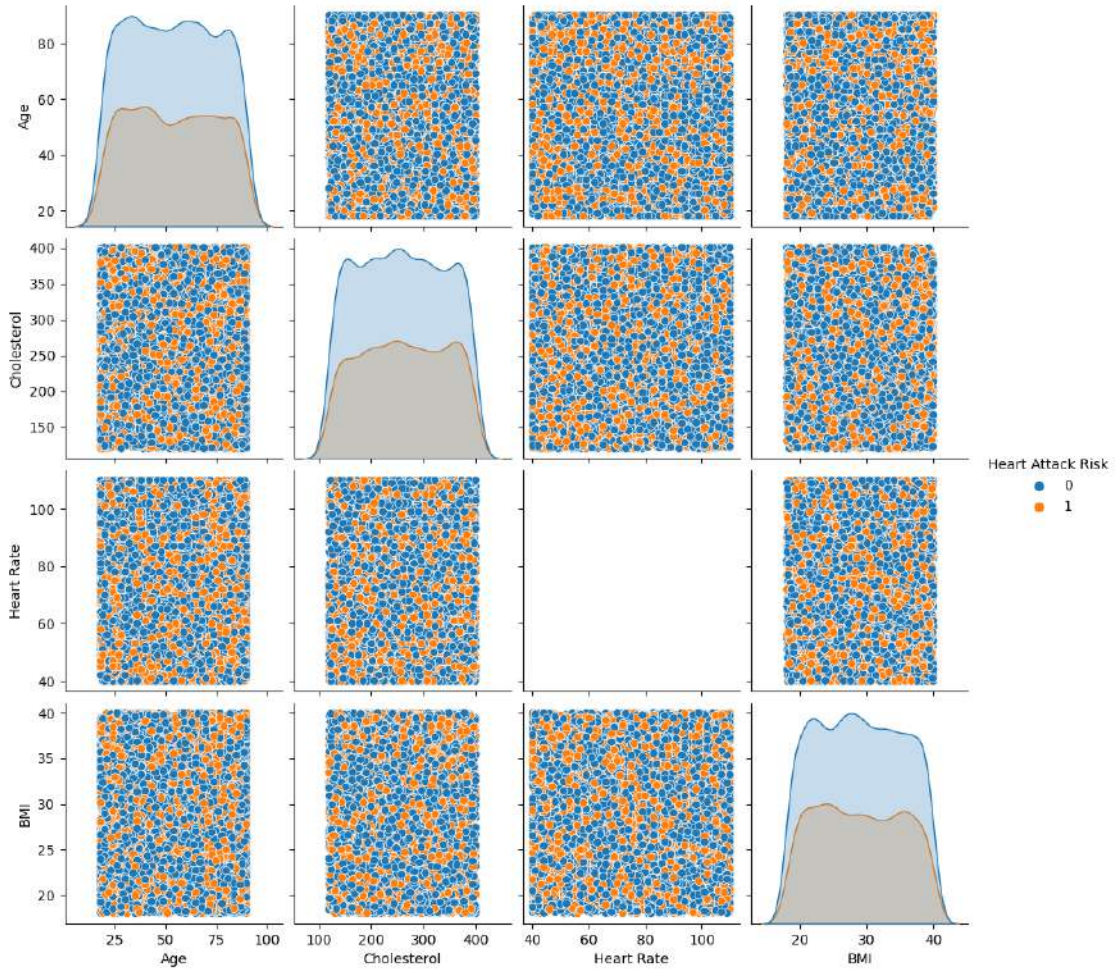
```
df_data_1.isnull().sum()
```

```
Patient ID                        0
Age                               0
Sex                               0
Cholesterol                       0
Blood Pressure                    0
Heart Rate                        0
Diabetes                          0
Family History                    0
Smoking                           0
Obesity                           0
Alcohol Consumption               0
Exercise Hours Per Week           0
Diet                              0
Previous Heart Problems           0
Medication Use                    0
Stress Level                      0
Sedentary Hours Per Day           0
Income                            0
BMI                               0
Triglycerides                     0
Physical Activity Days Per Week   0
Sleep Hours Per Day               0
Country                           0
Continent                         0
Hemisphere                        0
```

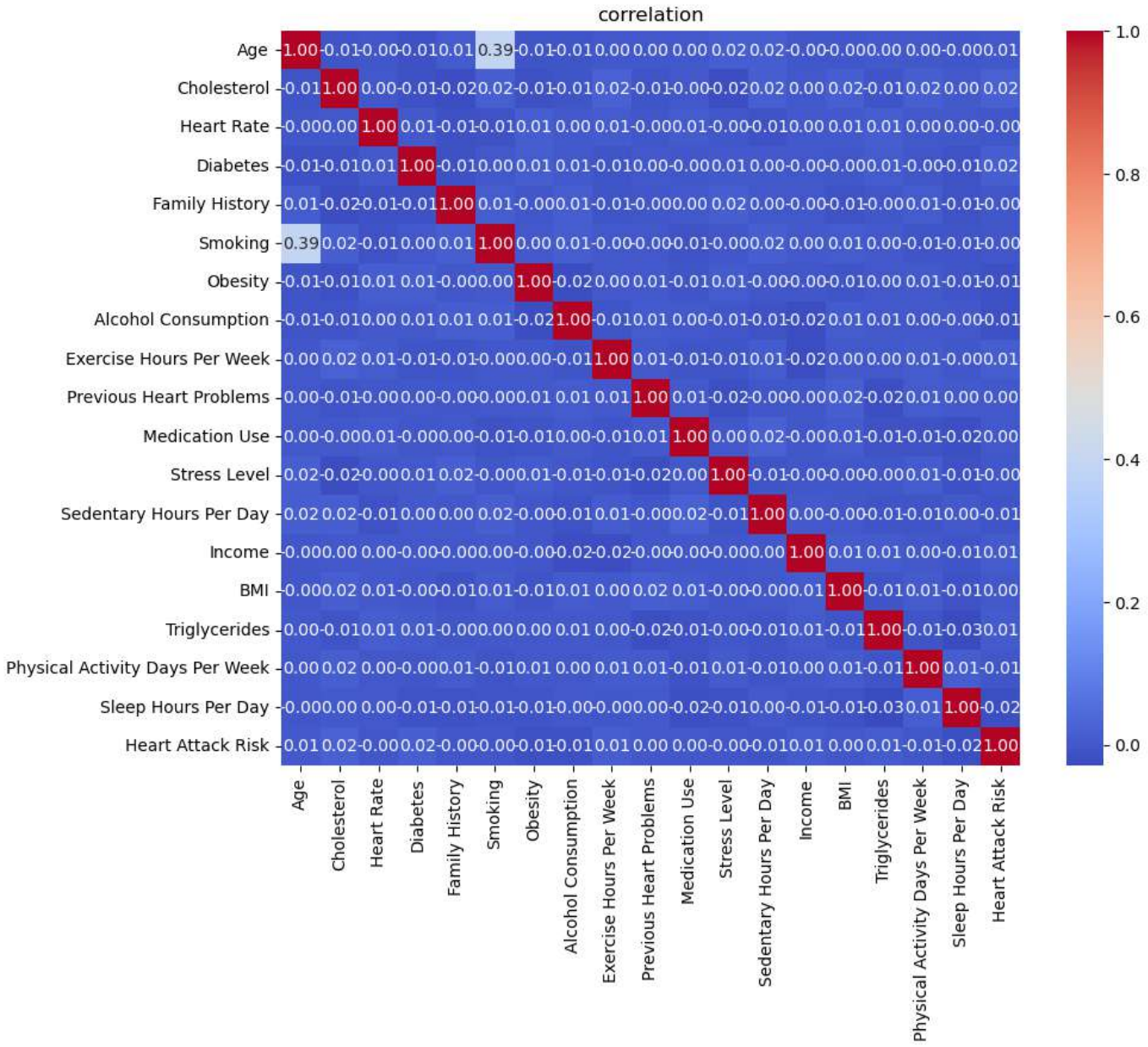
```
Heart Attack Risk                   0
dtype: int64
```

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize = (10,6))
sns.pairplot(df, vars=['Age', 'Cholesterol', 'Heart Rate', 'BMI'], hue='Heart Attack Risk')
plt.show()
```

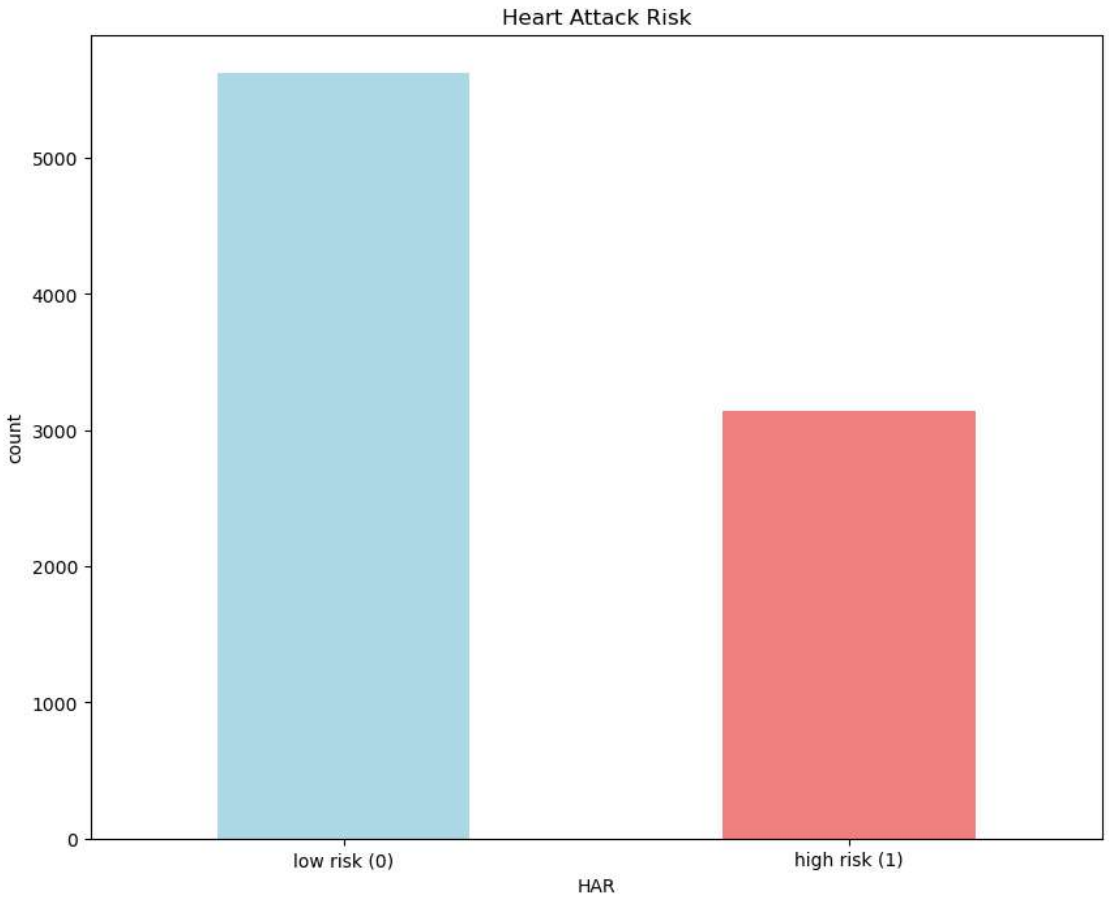
```
<Figure size 1000x600 with 0 Axes>
```




```
numeric_data =df.select_dtypes(include=["int64","float64"])
correlation_matrix = numeric_data.corr()
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix, annot = True, cmap = 'coolwarm', fmt = ".2f")
plt.title("correlation")
plt.show()
```




```
class_counts = df["Heart Attack Risk"].value_counts()
plt.figure(figsize =(10,8))
class_counts.plot(kind = "bar", color = ["lightblue","lightcoral"])
plt.title("Heart Attack Risk")
plt.xlabel("HAR")
plt.ylabel("count")
plt.xticks([0,1], labels = ["low risk (0)", "high risk (1)"], rotation = 0)
plt.show()
```




```
for column in df.columns:
    unique_values = df[column].unique()
    print(f"unique values for column {column}:\n{unique_values}\n")
```

```
unique values for column Patient ID:
['BMW7812' 'CZE1114' 'BNI9906' ... 'XKA5925' 'EPE6801' 'ZWN9666']

unique values for column Age:
[67 21 84 66 54 90 20 43 73 71 77 60 88 69 38 50 45 36 48 40 79 63 27 25
 86 42 52 29 30 47 44 33 51 70 85 31 56 24 74 72 55 26 53 46 57 22 35 39
 80 65 83 82 28 19 75 18 34 37 89 32 49 23 59 62 64 61 76 41 87 81 58 78
 68]

unique values for column Sex:
['Male' 'Female']

unique values for column Cholesterol:
[208 389 324 383 318 297 358 220 145 248 373 374 228 259 122 379 166 303
 340 294 359 202 133 159 271 273 328 154 135 197 321 375 360 263 201 347
 129 229 251 121 190 185 279 336 192 180 203 368 222 243 218 120 285 377
 369 311 139 266 153 339 329 333 398 124 183 163 362 390 200 396 255 209
 247 250 227 246 223 330 195 194 178 155 240 237 216 276 224 326 198 301
 314 304 334 213 254 230 316 277 388 206 384 205 261 308 338 382 291 168
```

```
 171 378 253 245 226 281 123 173 231 234 268 306 186 293 161 380 239 149
 320 219 335 265 126 307 270 225 193 148 296 136 364 353 252 232 387 299
 357 214 370 345 351 344 152 150 131 272 302 337 170 356 274 188 125 138
 376 181 184 275 394 128 217 399 283 289 284 327 262 212 350 385 162 141
 361 244 295 287 144 354 363 352 140 196 172 319 325 331 392 147 187 346
 286 151 300 165 343 366 317 386 158 157 242 241 365 257 348 175 298 269
 267 397 310 341 204 127 290 280 132 322 179 199 143 312 288 395 189 156
 238 381 391 355 210 400 260 235 167 256 249 207 130 134 137 305 236 315
 292 323 146 258 332 372 142 309 177 367 371 211 282 342 264 176 160 233
 313 164 349 221 191 174 393 278 215 169 182]

unique values for column Blood Pressure:
['158/88' '165/93' '174/99' ... '137/94' '94/76' '119/67']

unique values for column Heart Rate:
[ 72  98  73  93  48  84 107  68  55  97  70  85 102  40  56 104  71  69
  66  81  52 105  96  74  49  45  50  46  44 106  83  86  65 101  51  43
  79  90  94  78  92  54 109  61  64  82 110  42  63  41 100  76  75  58
  53  60  77  47  59  57  87  67  88  99  80  95 108  89  62 103  91]

unique values for column Diabetes:
[0 1]

unique values for column Family History:
[0 1]

unique values for column Smoking:
[1 0]

unique values for column Obesity:
[0 1]

unique values for column Alcohol Consumption:
[0 1]

unique values for column Exercise Hours Per Week:
[ 4.16818884  1.81324162  2.07835299 ...  3.14843791  3.78994983
 18.08174797]

unique values for column Diet:
['Average' 'Unhealthy' 'Healthy']

unique values for column Previous Heart Problems:
[0 1]

unique values for column Medication Use:
[0 1]

unique values for column Stress Level:
```

```
[ 9  1  6  2  7  4  5  8 10  3]

unique values for column Sedentary Hours Per Day:
[6.61500145 4.96345884 9.46342584 ... 2.37521373 0.02910426 9.00523438]

unique values for column Income:
[261404 285768 235282 ...  36998 209943 247338]

unique values for column BMI:
[31.25123273 27.19497335 28.17657068 ... 35.40614616 27.29402009
 32.91415086]

unique values for column Triglycerides:
[286 235 587 378 231 795 284 370 790 232 469 523 590 506 635 773  68 402
 517 247 747 360 358 526 605 667 316 551 482 718 297 661 558 209 586 743
 411 785 697 519 595 452 158 679 675 792 584 366 741 474  92 410 398 493
 614 682 106 216 408 628 481  67  82 305 164 211 511 766 547 327 367 681
 131  42 692 664 543 689 569 458 683 779 136 643 653  55 275 314 760 404
 576 690 648 385 255 468 784 509 205 109 530 654 331 485 250 113 377 180
 229 602 285 471 554 344 416 445 709 426 528 388 441 306 749 347 341 451
 356 336 455 223 262 239 555 363 489 788 121 553 617 174 167 563 665  65
 657 237 141 767 292 214 221 447 634 460 711  97 267 695 717 383 332 449
 701 524 549  31 276 744 128  52 394  54 739 407 751 436 473 218 129 579
 492 696 202 197 521 325  35 123 694 434 248 348 750 431 714 649 668 401
 610 244 691  88 532 777 420 350 652 413 754 753 457 122 312 778 676 775
 183 601 317 592 191  83  32 453 423 234 650 565 798 769 412  63 198  93
 764 737  94 298 288 735 190 281 146 574 359 155 719 466 273 515 187 544
 103 132 118 115  85  38 117 362 133 498 645 339 787 733 663 291 502  78
  81 257 624  91 374 270 797 446 464 450 722 556 184 428 796 656 134 196
 623 522 376 730 463  99 593  47 148 302  57 280 389 629 294 186 700 774
 181 375 467 603 616 380 495 698 318 207 780  51  84 425 310 126  56 472
 669 688 655  39 333 501 479 540 433 179 490 204 644 525 546 486 320 319
  58 591 165 732 195 478 461 631 301  50 315 194 199 160 149 527 406 161
 125 200 277 308  69 427 236  77 500 269  79 168 575 606 355 636  64 251
 245 228 287 800 483 791 260 604 536 559 124 254 159  73 542 390 755  60
  61 491  40 437 215 440 379 789 266 505 243 783 403 637 156 729 438 507
 725 562 324  87 253 626 541 364 456  30 182 621 494 776 442 429 684 219
  70  98 166  95 135 646 337 226 710 608 208 724 704 512 206 224 622 598
 465 119 293 630 386 513  45 578 261 217 715 282 391 580 192 399 249 396
 278 448 782 419 503 220  49 304 157 150 545 627 582 178 263  33 299 303
  66 763 256 139 651 756 372 345  48  46 421  43 771 210 781  41 508 353
 566 726 736 326 759 477 369 188 104 329 309 384 599 415 770 571 552 145
 632 373  71 550 583 322 475 357 673 454 757 201 100 274 258 613 233 330
 731 761 296 573 335 716 642 142 674 572 638 222 752 740 397 594 705 381
 615 539 242 499 435 680 535 238 283  89 589 666 678  76 176 620  75 721
 143 723 570  44 203 259 677 734 662 707 745 487 577 443 120 111 365 116
 538 162 742 212 581 313  36 400 619 609 252 706 264 290 138 300 346 712
  34 387 140 154 758 462 672 713  86 414 699 529 382 432 368 193  72 537
 560 189 342 531 311 241 685 497 640 321 480 144 585 171 727 660 799 600
```

```
 597 213 708 151 265 618 658 746 307  53 514 611 153 352 225 567 702 520
 417 102 607 548 647 476 762 147 424 459 409  74 510  37 323 240 175 786
 080 439 504 772 670  59 334 703 392  90 496 422 279 343 671 794 163 328
 625 272 227 152 105 693  96 484 568 633 659 230 112 793 101 172 110 612
 185 289 418 533 686 641 169 349 173 516  62 557 596 728 371 738 444 561
 114 765 338 588 246 295 564 488 177 687 395 518 127 639 137 354 271 107
 340 534 768 130 720 405 430 268 108 748 351 393 361 170 470]

unique values for column Physical Activity Days Per Week:
[0 1 4 3 5 6 7 2]

unique values for column Sleep Hours Per Day:
[ 6  7  4  5 10  8  9]

unique values for column Country:
['Argentina' 'Canada' 'France' 'Thailand' 'Germany' 'Japan' 'Brazil'
 'South Africa' 'United States' 'Vietnam' 'China' 'Italy' 'Spain' 'India'
 'Nigeria' 'New Zealand' 'South Korea' 'Australia' 'Colombia'
 'United Kingdom']

unique values for column Continent:
['South America' 'North America' 'Europe' 'Asia' 'Africa' 'Australia']

unique values for column Hemisphere:
['Southern Hemisphere' 'Northern Hemisphere']

unique values for column Heart Attack Risk:
[0 1]
```

```python
from sklearn.preprocessing import LabelEncoder
x=[]
lab=LabelEncoder()
for i in df.select_dtypes(include='object').columns.values:
    df[i]=lab.fit_transform(df[i])
for i in df.columns.values:
    print(df[i].value_counts())
    print()
```

```
521     1
1038    1
1601    1
7555    1
4141    1
       ..
4539    1
1663    1
1538    1
4056    1
8719    1
```

```
Name: Patient ID, Length: 8763, dtype: int64

90    152
42    150
33    147
59    147
29    137
     ...
75    102
72    101
39    100
47     99
51     82
Name: Age, Length: 73, dtype: int64

1    6111
0    2652
Name: Sex, dtype: int64

235    52
360    47
149    46
218    46
251    45
       ..
248    20
186    20
328    20
398    20
397    19
Name: Cholesterol, Length: 281, dtype: int64

2005    8
87      8
283     7
98      7
3295    7
       ..
2346    1
2318    1
3395    1
3863    1
838     1
Name: Blood Pressure, Length: 3915, dtype: int64

94     157
97     146
57     143
52     140
```

```
104    139
       ...
70     107
48     107
79     105
96      97
73      93
Name: Heart Rate, Length: 71, dtype: int64

1    5716
0    3047
Name: Diabetes, dtype: int64

0    4443
1    4320
Name: Family History, dtype: int64

1    7859
0     904
Name: Smoking, dtype: int64

1    4394
0    4369
Name: Obesity, dtype: int64

1    5241
0    3522
Name: Alcohol Consumption, dtype: int64

4.168189     1
18.477430    1
11.883523    1
19.353157    1
19.365546    1
            ..
9.884039     1
12.644947    1
1.089868     1
10.500477    1
18.081748    1
Name: Exercise Hours Per Week, Length: 8763, dtype: int64

1    2960
0    2912
2    2891
Name: Diet, dtype: int64

0    4418
1    4345
```

```
Name: Previous Heart Problems, dtype: int64


0    4396
1    4367
Name: Medication Use, dtype: int64


2     913
4     910
7     903
9     887
8     879
3     868
1     865
5     860
6     855
10    823
Name: Stress Level, dtype: int64


6.615001     1
0.772688     1
0.723868     1
10.125510    1
2.054331     1
            ..
11.921800    1
0.087028     1
9.198925     1
3.383760     1
9.005234     1
Name: Sedentary Hours Per Day, Length: 8763, dtype: int64


225278    4
194461    3
195282    3
220507    2
139451    2
         ..
44744     1
85563     1
20443     1
258704    1
247338    1
Name: Income, Length: 8615, dtype: int64


31.251233    1
39.385227    1
36.280438    1
18.218558    1
23.885840    1
```

```
           ..

28.358868    1
22.539845    1
34.721372    1
18.881817    1
32.914151    1
Name: BMI, Length: 8763, dtype: int64


799    25
507    22
121    22
593    22
469    22
       ..
120     3
213     3
185     3
295     3
130     2
Name: Triglycerides, Length: 771, dtype: int64


3    1143
1    1121
2    1109
7    1095
5    1079
4    1077
6    1074
0    1065
Name: Physical Activity Days Per Week, dtype: int64


10    1293
8     1288
6     1276
7     1270
5     1263
9     1192
4     1181
Name: Sleep Hours Per Day, dtype: int64


7     477
0     471
2     462
17    457
1     449
12    448
6     446
3     440
4     436
```

```
 11    435
10    433
9     431
15    430
5     429
16    428
13    425
19    425
18    420
8     412
14    409
Name: Country, dtype: int64

1    2543
3    2241
5    1362
2     884
0     873
4     860
Name: Continent, dtype: int64

0    5660
1    3103
Name: Hemisphere, dtype: int64

0    5624
1    3139
Name: Heart Attack Risk, dtype: int64
```

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
# Features (X) and target (y)
features = df.drop(columns=['Patient ID', 'Heart Attack Risk'])
target = df['Heart Attack Risk']
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2,
random_state=42)
# Initialize and train the RandomForestClassifier
classifier = RandomForestClassifier(random_state=42)
classifier.fit(X_train, y_train)
# Make predictions on the test set
predictions = classifier.predict(X_test)
# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.6400456360524814
```

4.Documentation & Reporting

Documentation is crucial to capture the project's journey and make the insights accessible. We will create a comprehensive report encompassing the following sections.

4.1. Project Outline

This section will provide an overview of the project's objective and approach. It will outline the structure of the report and what readers can expect to find.

4.2. Business Insights

The project's ultimate goal is to derive actionable business insights. This section will explain how the analysis findings are translated into valuable recommendations for decision-making and strategy development.

4.3. Dataset Description

Detailed information on the selected datasets, including the reasons for their selection and how they were integrated into IBM Cloud Databases.

4.4. Database Setup and Data Exploration

A comprehensive explanation of the database setup and data exploration process. It will shed light on the infrastructure and methods employed to explore the datasets.

4.5. Analysis Techniques and Visualization

This section will delve into the analysis techniques used and the methods of visualization applied to convey the results effectively.

5.Conclusion:

In this part we will document our project and prepare it for submission.Document the big data analysis project and prepare it for submission.

Documentation:

Outline the project's objective, design thinking process, and development phases.Describe the selected dataset, database setup, analysis techniques, and visualization methods used.Explained how the analysis findings translate into valuable business insights.

Remember to tailor this structure to your specific project and organization's needs. Each section should provide a clear and concise overview of the project's details and its value to the business.