Name: Harshit Hemant Gupta

IU ID: 2001096665

Subject: FA24-BL-ENGR-E534 BIG DATA APPLICATIONS

Mini Project: Fall24

1. Dataset Selection (0 Points)

● **Task:** Choose a dataset of your own choice that is large enough to derive meaningful insights.

○ **Examples:** Financial transactions, customer behavior, IoT sensor data.

○ Ensure the dataset contains diverse columns to allow for cleaning, transformation, and aggregation.

California Independent Medical Review

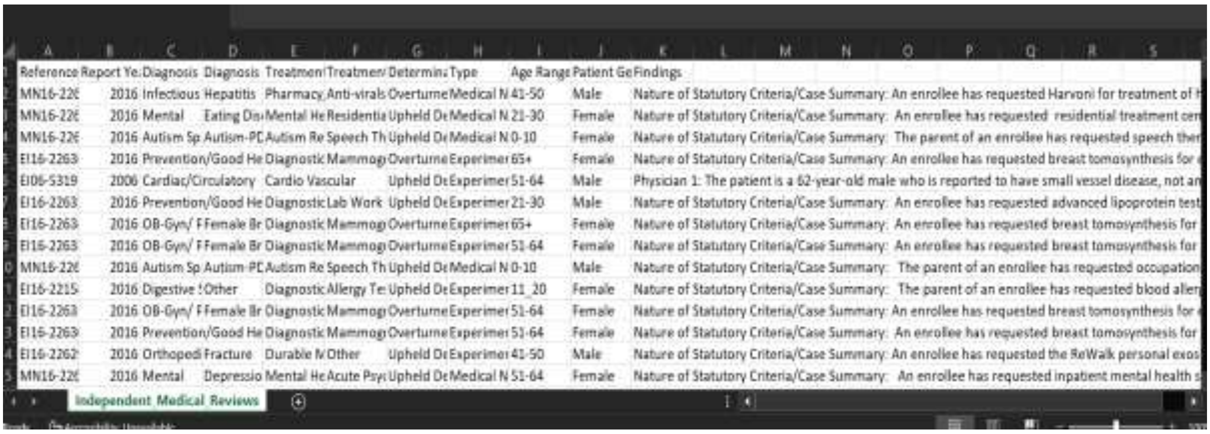
https://www.kaggle.com/datasets/prasad22/ca-independent-medical-review/data

Description:

This dataset, sourced from the California Department of Managed Health Care (DMHC), includes all decisions from Independent Medical Reviews (IMRs) conducted by the DMHC since January 1, 2001. IMRs are impartial evaluations of health care services that have been denied, delayed, or modified by a health plan, often on grounds of being deemed unnecessary, experimental, or non-urgent. When an IMR decision favors the enrollee, the health plan is required to approve the requested treatment or service.

File Information:

- File Format: CSV
- Number of Rows: 19246
- Number of Columns: 11



Reference	Report Ye	Diagnosis	Diagnosis	Treatmen	Treatmen	Determin	Type	Age Range	Patient Ge	Findings
MN16-226	2016	Infectious	Hepatitis	Pharmacy	Anti-virals	Overturne	Medical N	41-50	Male	Nature of Statutory Criteria/Case Summary: An enrollee has requested Harvoni for treatment of H
MN16-226	2016	Mental	Eating Dis	Mental He	Residentia	Upheld De	Medical N	21-30	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested residential treatment cen
MN16-226	2016	Autism Sp	Autism-PC	Autism Re	Speech Th	Upheld De	Medical N	0-10	Female	Nature of Statutory Criteria/Case Summary: The parent of an enrollee has requested speech ther
EI16-2263	2016	Prevention/Good He		Diagnostic	Mammogr	Overturne	Experimen	65+	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested breast tomosynthesis for e
EI06-5319	2006	Cardiac/Circulatory		Cardio Vascular		Upheld De	Experimen	51-64	Male	Physician 1: The patient is a 62-year-old male who is reported to have small vessel disease, not an
EI16-2263	2016	Prevention/Good He		Diagnostic	Lab Work	Upheld De	Experimen	21-30	Male	Nature of Statutory Criteria/Case Summary: An enrollee has requested advanced lipoprotein test
EI16-2263	2016	OB-Gyn/ F	Female Br	Diagnostic	Mammogr	Overturne	Experimen	65+	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested breast tomosynthesis for
EI16-2263	2016	OB-Gyn/ F	Female Br	Diagnostic	Mammogr	Overturne	Experimen	51-64	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested breast tomosynthesis for
MN16-226	2016	Autism Sp	Autism-PC	Autism Re	Speech Th	Upheld De	Medical N	0-10	Male	Nature of Statutory Criteria/Case Summary: The parent of an enrollee has requested occupation
EI16-2215	2016	Digestive S	Other	Diagnostic	Allergy Te	Upheld De	Experimen	11_20	Female	Nature of Statutory Criteria/Case Summary: The parent of an enrollee has requested blood aller
EI16-2263	2016	OB-Gyn/ F	Female Br	Diagnostic	Mammogr	Overturne	Experimen	51-64	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested breast tomosynthesis for e
EI16-2263	2016	Prevention/Good He		Diagnostic	Mammogr	Overturne	Experimen	51-64	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested breast tomosynthesis for
EI16-2262	2016	Orthopedi	Fracture	Durable M	Other	Upheld De	Experimen	41-50	Male	Nature of Statutory Criteria/Case Summary: An enrollee has requested the ReWalk personal exos
MN16-226	2016	Mental	Depressio	Mental He	Acute Psy	Upheld De	Medical N	51-64	Female	Nature of Statutory Criteria/Case Summary: An enrollee has requested inpatient mental health s

Data Schema:

Field Name	Description	Data type
Reference ID	Unique identifier for the case	Plain Text
Report Year	Year the case was reported	Number
Diagnosis Category	The main diagnosis category	Plain Text
Diagnosis Sub Category	The secondary diagnosis category	Plain Text
Treatment Category	The main treatment category	Plain Text
Treatment Sub Category	The secondary treatment category	Plain Text
Determination	Indicates if the determination was upheld or overturned	Plain Text
Type	Indicates the type of case (Experimental/Investigational, Urgent Care, Medical Necessity)	Plain Text
Age Range	The age of the patient	Number
Patient Gender	The gender of the patient	Plain Text
Findings	A summary of the case findings	Plain Text

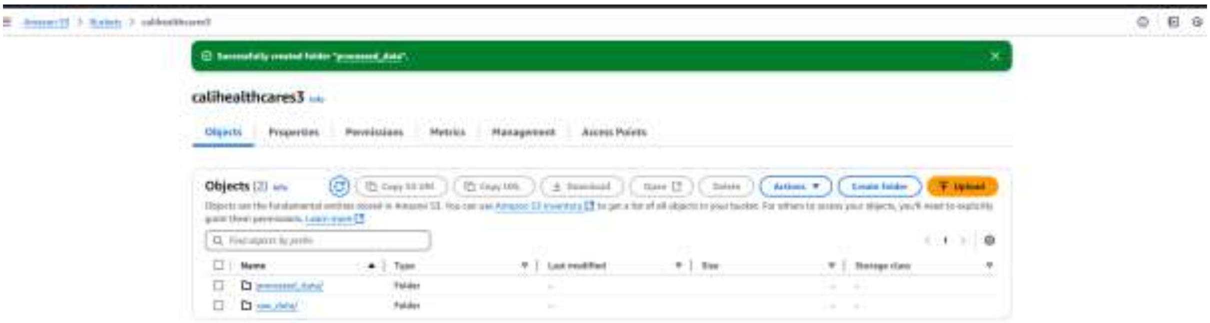
2. Environment Setup (2.5 points)

□ **AWS S3 for Data Storage**: Demonstrates the creation and configuration of an AWS S3 bucket, showcasing the process of uploading raw data into S3 to serve as a centralized storage for further data processing.

□ **Linux Environment with PySpark**: Highlights the setup of a Linux-based environment (e.g., an AWS EC2 instance), installation of PySpark, and configuration of AWS CLI to enable seamless interaction with the S3 bucket for distributed data processing tasks.

1. AWS S3 for Data Storage (1 Point)

a. Step 1: Create an S3 bucket to store both raw and processed data.



Successfully created folder "processed_data".

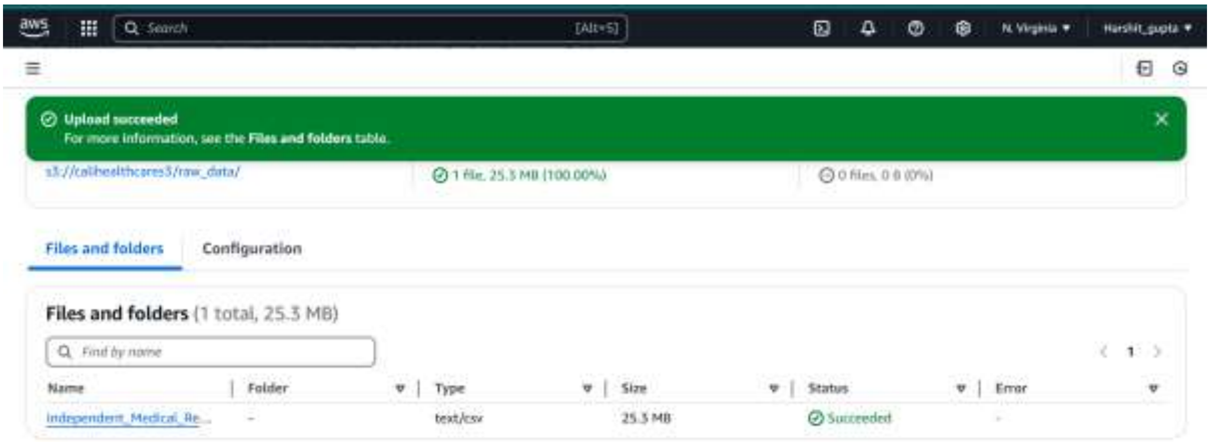
calihealthcares3 Info

Objects (2) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	processed_data/	Folder	-	-	-
□	raw_data/	Folder	-	-	-

b. Step 2: Upload the raw dataset to the S3 bucket.



Upload succeeded
For more information, see the **Files and folders** table.

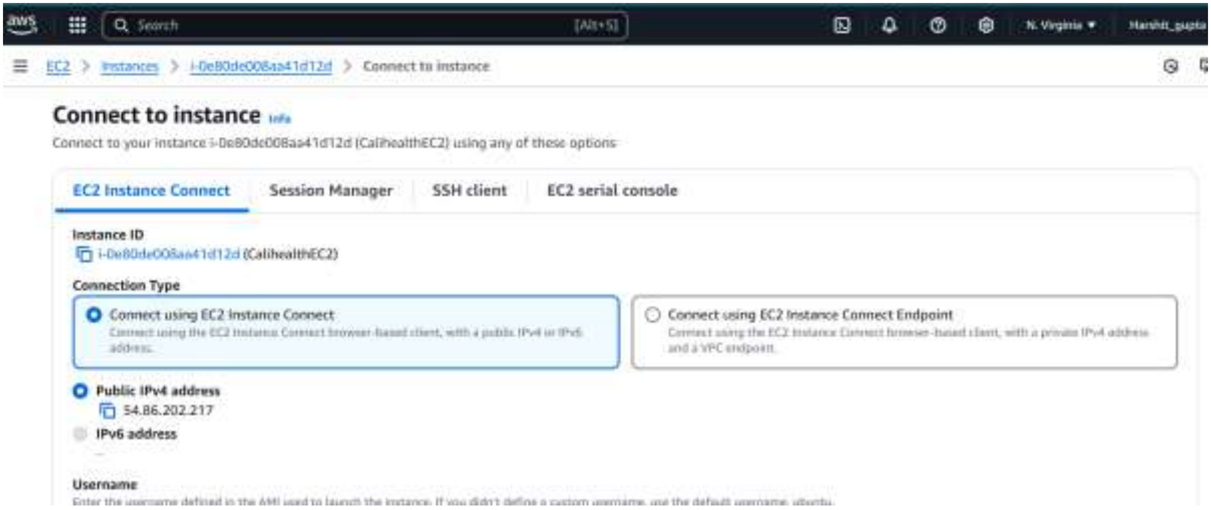
s3://calihealthcares3/raw_data/ ✓ 1 file, 25.3 MB (100.00%) ○ 0 files, 0 B (0%)

Files and folders (1 total, 25.3 MB)

Name	Folder	Type	Size	Status	Error
Independent_Medical_Re...	-	text/csv	25.3 MB	✓ Succeeded	-

2. Linux Environment with PySpark (1.5 Point)

a. Step 1: Set up a Linux-based environment, either locally or using an AWS EC2 instance. **Recommendation:** Use an AWS EC2 instance for better scalability and AWS integration.



Connect to instance Info

Connect to your instance i-0e80de008aa41d12d (CalihealthEC2) using any of these options

Instance ID
i-0e80de008aa41d12d (CalihealthEC2)

Connection Type

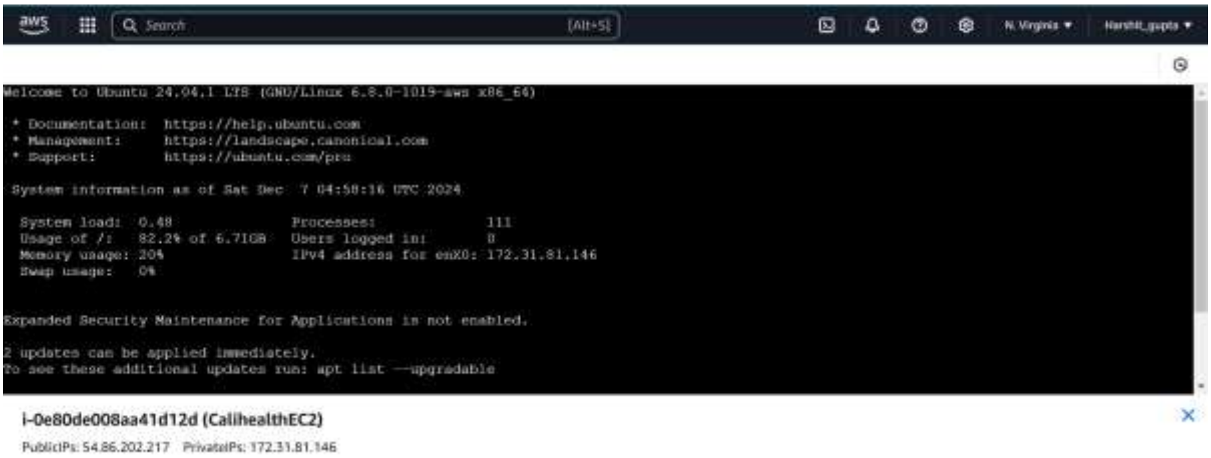
● Connect using EC2 Instance Connect
Connect using the EC2 Instance Connect browser-based client, with a public IPv4 or IPv6 address.

○ Connect using EC2 Instance Connect Endpoint
Connect using the EC2 Instance Connect browser-based client, with a private IPv4 address and a VPC endpoint.

● Public IPv4 address
54.86.202.217

○ IPv6 address

Username
Enter the username defined in the AMI used to launch the instance. If you didn't define a custom username, use the default username, ubuntu.

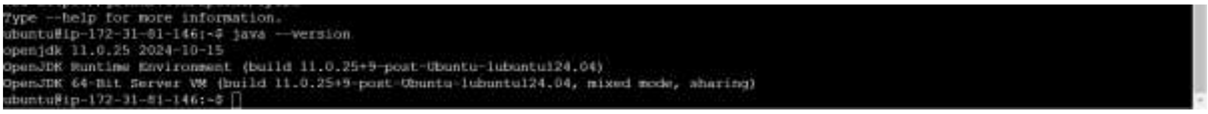


```
Welcome to Ubuntu 24.04.1 LTS (GNU/Linux 6.8.0-1019-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 System information as of Sat Dec  7 04:58:16 UTC 2024

  System load:  0.48               Processes:             111
  Usage of /:   82.2% of 6.71GB    Users logged in:       0
  Memory usage: 20%                IPv4 address for enX0: 172.31.81.146
  Swap usage:   0%


Expanded Security Maintenance for Applications is not enabled.

2 updates can be applied immediately.
To see these additional updates run: apt list --upgradable
```

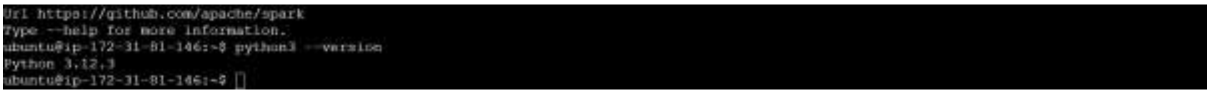
i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.86.202.217   PrivateIPs: 172.31.81.146

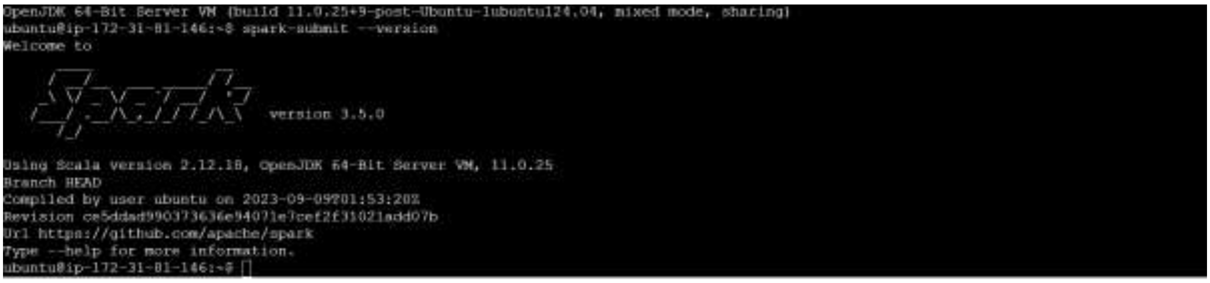
b. Step 2: Install PySpark for distributed data processing.



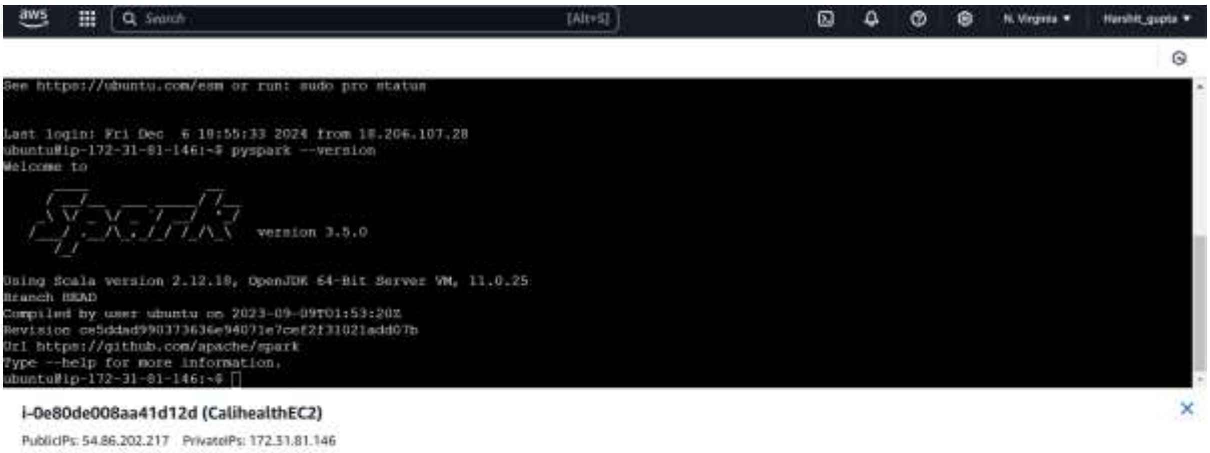
```
Type --help for more information.
ubuntu@ip-172-31-81-146:~$ java --version
openjdk 11.0.25 2024-10-15
OpenJDK Runtime Environment (build 11.0.25+9-post-Ubuntu-1ubuntu124.04)
OpenJDK 64-Bit Server VM (build 11.0.25+9-post-Ubuntu-1ubuntu124.04, mixed mode, sharing)
ubuntu@ip-172-31-81-146:~$ 
```



```
Url https://github.com/apache/spark
Type --help for more information.
ubuntu@ip-172-31-81-146:~$ python3 --version
Python 3.12.3
ubuntu@ip-172-31-81-146:~$ 
```



```
OpenJDK 64-Bit Server VM (build 11.0.25+9-post-Ubuntu-1ubuntu124.04, mixed mode, sharing)
ubuntu@ip-172-31-81-146:~$ spark-submit --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 11.0.25
Branch HEAD
Compiled by user ubuntu on 2023-09-09T01:53:20Z
Revision ce5ddad990373636e94071e7cef2f31021add07b
Url https://github.com/apache/spark
Type --help for more information.
ubuntu@ip-172-31-81-146:~$
```

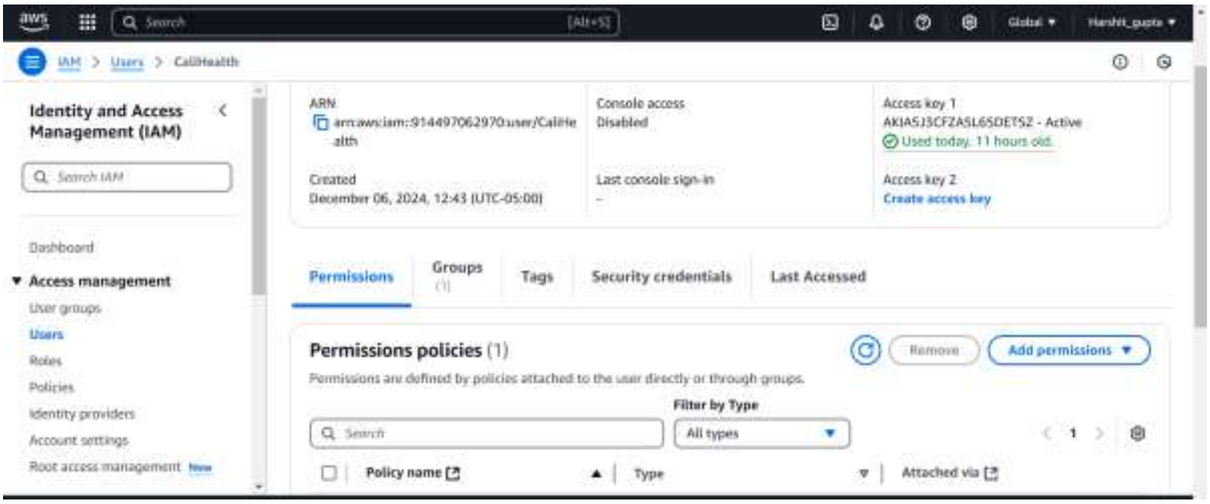


```
See https://ubuntu.com/esm or run: sudo pro status


Last login: Fri Dec  6 18:55:33 2024 from 18.206.107.28
ubuntu@ip-172-31-81-146:~$ pyspark --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 11.0.25
Branch HEAD
Compiled by user ubuntu on 2023-09-09T01:53:20Z
Revision ce5ddad990373636e94071e7cef2f31021add07b
Url https://github.com/apache/spark
Type --help for more information.
ubuntu@ip-172-31-81-146:~$
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.86.202.217   PrivateIPs: 172.31.81.146

c. Step 3: Configure AWS CLI to interact with S3 buckets.



ARN
arn:aws:iam::914497062970:user/CaliHealth

Created
December 06, 2024, 12:43 (UTC-05:00)

Console access
Disabled

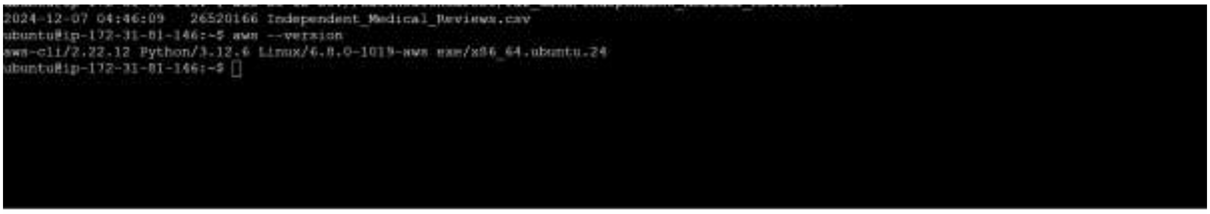
Last console sign-in
-

Access key 1
AKIA5J3CFZA5L65DETS2 - Active
Used today. 11 hours old.

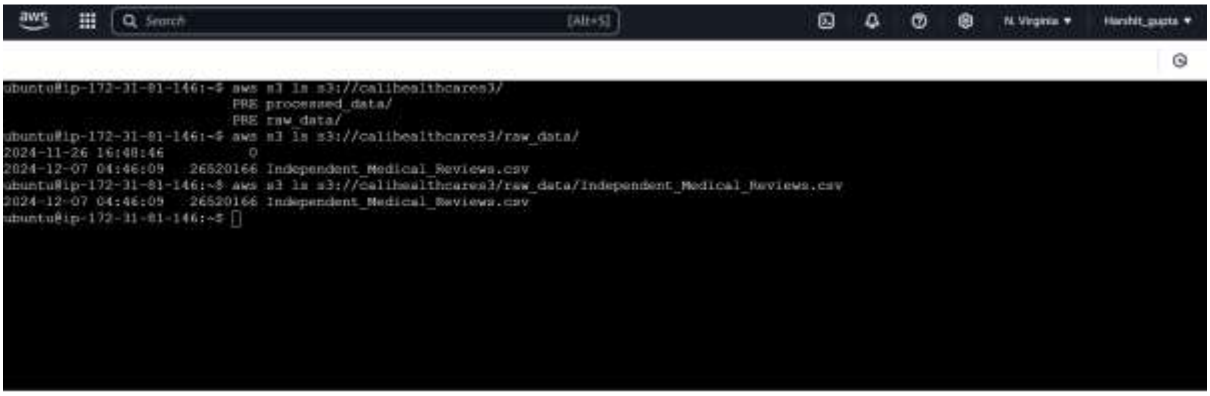
Access key 2
Create access key

Permissions policies (1)

Permissions are defined by policies attached to the user directly or through groups.



```
2024-12-07 04:46:09   26520166 Independent_Medical_Reviews.csv
ubuntu@ip-172-31-81-146:~$ aws --version
aws-cli/2.22.12 Python/3.12.6 Linux/6.8.0-1019-aws exe/x86_64.ubuntu.24
ubuntu@ip-172-31-81-146:~$
```



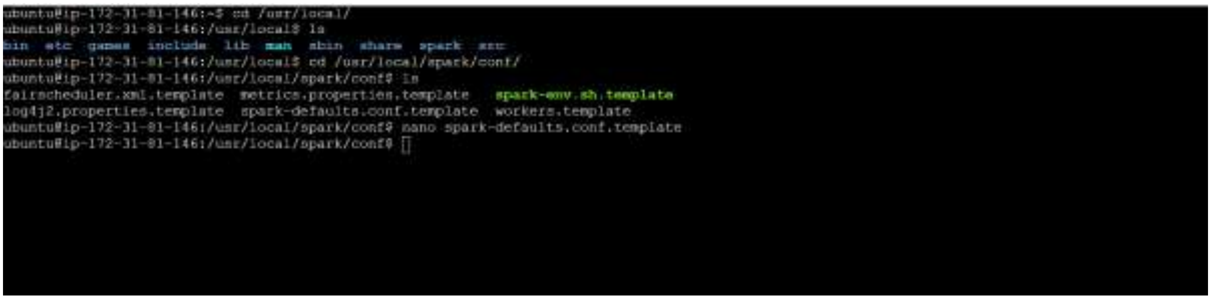
```
ubuntu@ip-172-31-81-146:~$ aws s3 ls s3://calihealthcares3/
                           PRE processed_data/
                           PRE raw_data/
ubuntu@ip-172-31-81-146:~$ aws s3 ls s3://calihealthcares3/raw_data/
2024-11-26 16:48:46          0
2024-12-07 04:46:09   26520166 Independent_Medical_Reviews.csv
ubuntu@ip-172-31-81-146:~$ aws s3 ls s3://calihealthcares3/raw_data/Independent_Medical_Reviews.csv
2024-12-07 04:46:09   26520166 Independent_Medical_Reviews.csv
ubuntu@ip-172-31-81-146:~$ 
```

3. Data Pipeline Tasks (6 points)

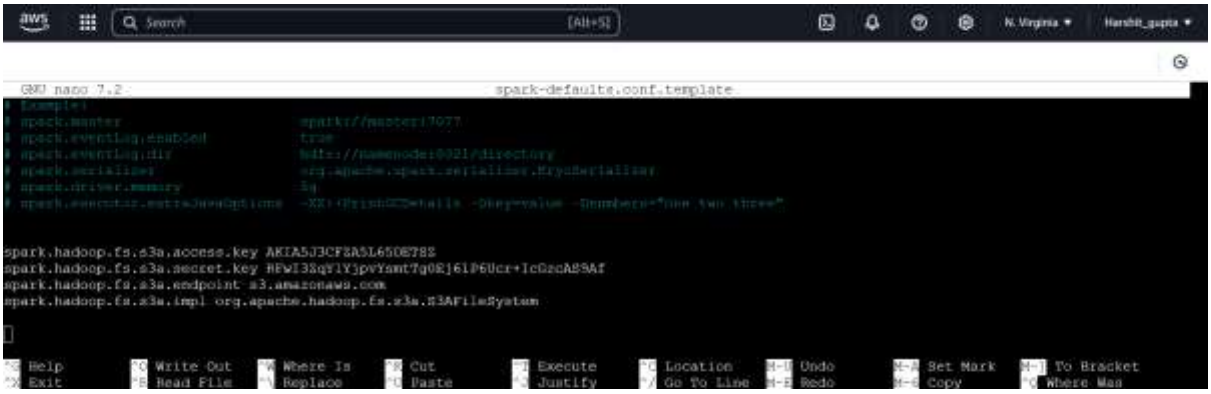
Task 1: Data Ingestion from S3 (1 Point)

● **Objective:** Pull raw data from S3 into the PySpark environment.

● **Steps:** 1. Use AWS CLI or PySpark's built-in S3 support to load the dataset directly.

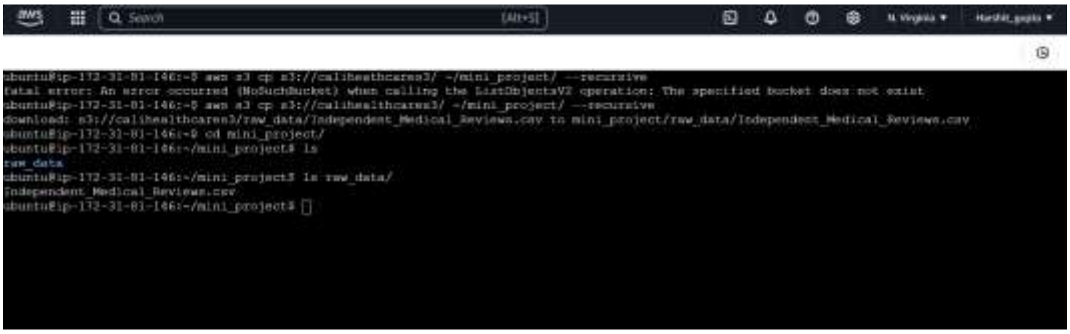


```
ubuntu@ip-172-31-81-146:~$ cd /usr/local/
ubuntu@ip-172-31-81-146:/usr/local$ ls
bin  etc  games  include  lib  man  sbin  share  spark  src
ubuntu@ip-172-31-81-146:/usr/local$ cd /usr/local/spark/conf/
ubuntu@ip-172-31-81-146:/usr/local/spark/conf$ ls
fairscheduler.xml.template  metrics.properties.template   spark-env.sh.template
log4j2.properties.template  spark-defaults.conf.template  workers.template
ubuntu@ip-172-31-81-146:/usr/local/spark/conf$ nano spark-defaults.conf.template
ubuntu@ip-172-31-81-146:/usr/local/spark/conf$ 
```

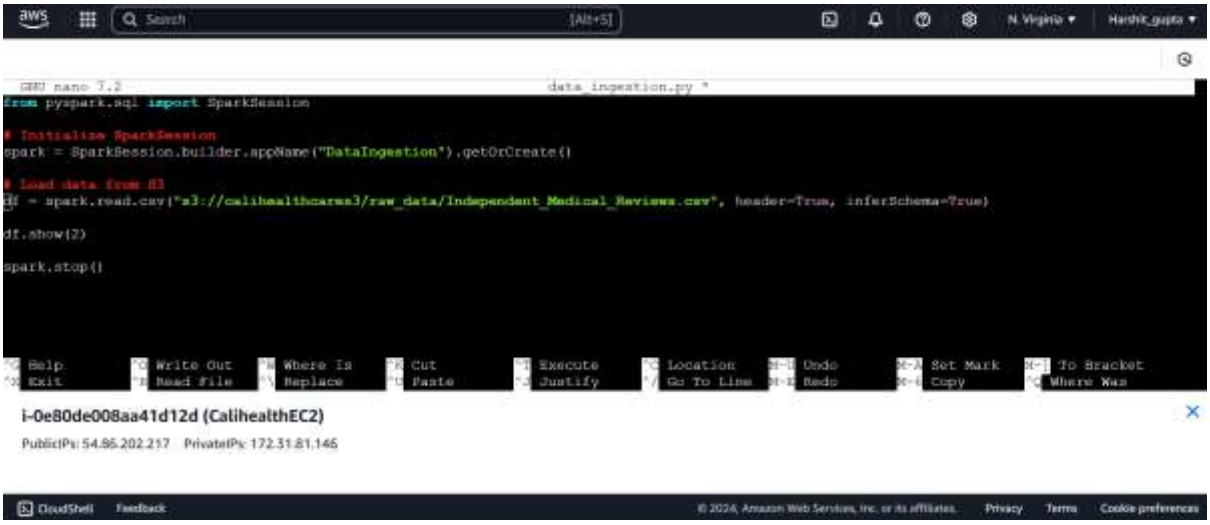


```
  GNU nano 7.2                                   spark-defaults.conf.template
# Example:
# spark.master                     spark://master:7077
# spark.eventLog.enabled           true
# spark.eventLog.dir               hdfs://namenode:8021/directory
# spark.serializer                 org.apache.spark.serializer.KryoSerializer
# spark.driver.memory              5g
# spark.executor.extraJavaOptions  -XX:+PrintGCDetails -Dkey=value -Dnumbers="one two three"

spark.hadoop.fs.s3a.access.key AKIA5J3CFZA5L65OE78Z
spark.hadoop.fs.s3a.secret.key HFwI3ZqYlYjpvYamtTg0Ej61P6Ucr+IcGzcAS9Af
spark.hadoop.fs.s3a.endpoint s3.amazonaws.com
spark.hadoop.fs.s3a.impl org.apache.hadoop.fs.s3a.S3AFileSystem

^G Help       ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location   M-U Undo      M-A Set Mark  M-] To Bracket
^X Exit       ^R Read File  ^\ Replace    ^U Paste      ^J Justify    ^/ Go To Line M-E Redo      M-6 Copy      ^Q Where Was
```

2. Confirm successful ingestion by inspecting the dataset.



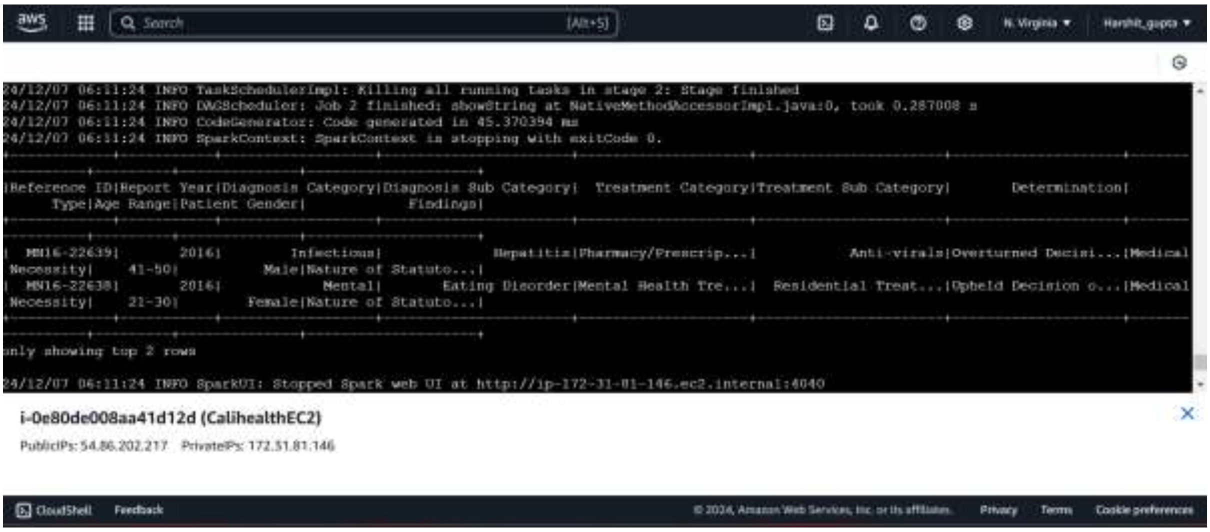
```
ubuntu@ip-172-31-81-146:~$ aws s3 cp s3://caliheathcares3/ ~/mini_project/ --recursive
fatal error: An error occurred (NoSuchBucket) when calling the ListObjectsV2 operation: The specified bucket does not exist
ubuntu@ip-172-31-81-146:~$ aws s3 cp s3://calihealthcares3/ ~/mini_project/ --recursive
download: s3://calihealthcares3/raw_data/Independent_Medical_Reviews.csv to mini_project/raw_data/Independent_Medical_Reviews.csv
ubuntu@ip-172-31-81-146:~$ cd mini_project/
ubuntu@ip-172-31-81-146:~/mini_project$ ls
raw_data
ubuntu@ip-172-31-81-146:~/mini_project$ ls raw_data/
Independent_Medical_Reviews.csv
ubuntu@ip-172-31-81-146:~/mini_project$ 
```



```
  GNU nano 7.2                                 data_ingestion.py *
from pyspark.sql import SparkSession

# Initialize SparkSession
spark = SparkSession.builder.appName("DataIngestion").getOrCreate()

# Load data from S3
df = spark.read.csv("s3://calihealthcares3/raw_data/Independent_Medical_Reviews.csv", header=True, inferSchema=True)

df.show(2)

spark.stop()
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.86.202.217 PrivateIPs: 172.31.81.146



```
24/12/07 06:11:24 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/12/07 06:11:24 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.287008 s
24/12/07 06:11:24 INFO CodeGenerator: Code generated in 45.370394 ms
24/12/07 06:11:24 INFO SparkContext: SparkContext is stopping with exitCode 0.
+------------+-----------+------------------+---------------------+--------------------+----------------------+---------------------+--------------
-----------+----------+--------------+----------------------+
|Reference ID|Report Year|Diagnosis Category|Diagnosis Sub Category|  Treatment Category|Treatment Sub Category|       Determination|
     Type|Age Range|Patient Gender|              Findings|
+------------+-----------+------------------+---------------------+--------------------+----------------------+---------------------+--------------
-----------+----------+--------------+----------------------+
|  MN16-22639|       2016|        Infectious|            Hepatitis|Pharmacy/Prescrip...|           Anti-virals|Overturned Decisi...|Medical
 Necessity|    41-50|          Male|Nature of Statuto...|
|  MN16-22638|       2016|            Mental|      Eating Disorder|Mental Health Tre...|  Residential Treat...|Upheld Decision o...|Medical
 Necessity|    21-30|        Female|Nature of Statuto...|
+------------+-----------+------------------+---------------------+--------------------+----------------------+---------------------+--------------
-----------+----------+--------------+----------------------+
only showing top 2 rows

24/12/07 06:11:24 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-81-146.ec2.internal:4040
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.86.202.217 PrivateIPs: 172.31.81.146

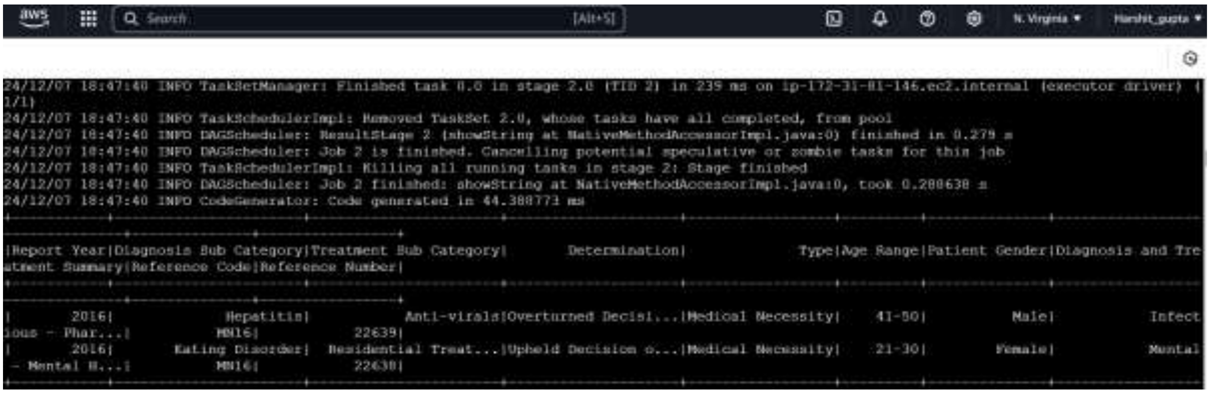
Image: Displays the raw data being pulled from an AWS S3 bucket into the PySpark environment.

Explanation: The image should show the use of either PySpark's spark.read functionality or the AWS CLI to load the dataset into the working environment. A preview of the successfully ingested data confirms that the process was completed without errors, ensuring all data fields and rows are intact.

Task 2: Data Processing with PySpark (2 Point)

All necessary techniques for cleaning data, such as transforming, aggregating, and removing outliers, should be applied.

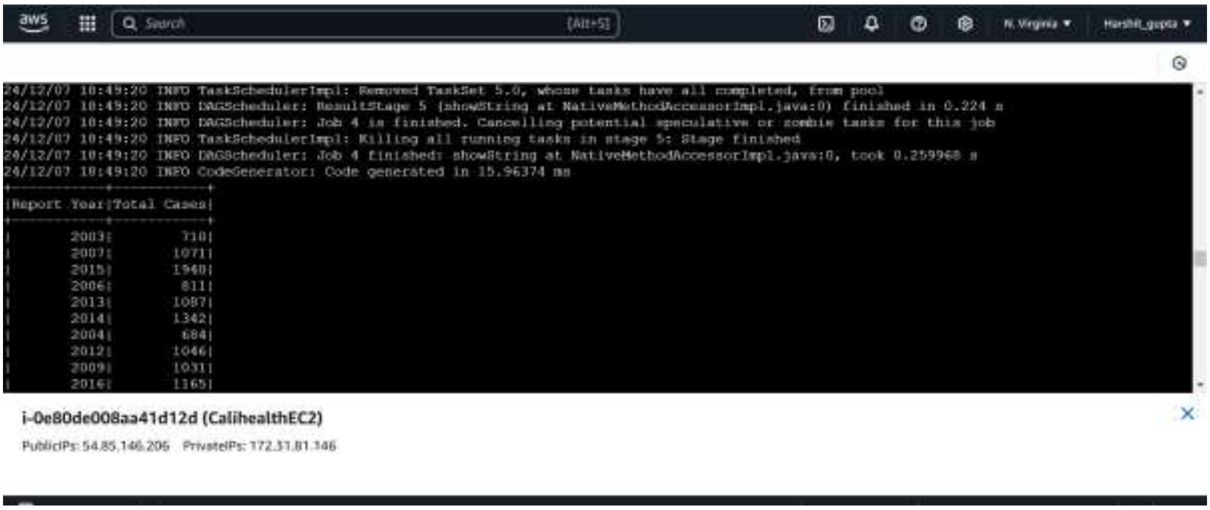
1. **Data Transformation:** Create at least 2 new columns (e.g., Year , Month) to aid in analysis.



```
24/12/07 18:47:40 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 239 ms on ip-172-31-81-146.ec2.internal (executor driver) (
1/1)
24/12/07 18:47:40 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
24/12/07 18:47:40 INFO DAGScheduler: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0) finished in 0.279 s
24/12/07 18:47:40 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 18:47:40 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/12/07 18:47:40 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.288638 s
24/12/07 18:47:40 INFO CodeGenerator: Code generated in 44.388773 ms
+-----------+-----------------------+-----------------------+--------------------+-----------------+---------+--------------+-----------------
-----------+--------------+----------------+
|Report Year|Diagnosis Sub Category|Treatment Sub Category|       Determination|             Type|Age Range|Patient Gender|Diagnosis and Tre
atment Summary|Reference Code|Reference Number|
+-----------+-----------------------+-----------------------+--------------------+-----------------+---------+--------------+-----------------
-----------+--------------+----------------+
|       2016|              Hepatitis|            Anti-virals|Overturned Decisi...|Medical Necessity|    41-50|          Male|           Infect
ious - Phar...|          MN16|           22639|
|       2016|         Eating Disorder|   Residential Treat...|Upheld Decision o...|Medical Necessity|    21-30|        Female|           Mental
 - Mental H...|          MN16|           22638|
+-----------+-----------------------+-----------------------+--------------------+-----------------+---------+--------------+-----------------
```

2. **Data Aggregation:** Compute at least 5 key metrics, such as Total revenue by region, Monthly spending trends, Top 10 customers by transaction value (These metrics may vary depending on the specifics of your dataset, but the goal is to aggregate the data in ways that enable meaningful analysis and decision-making.)

<u>Total cases by Report Year</u>

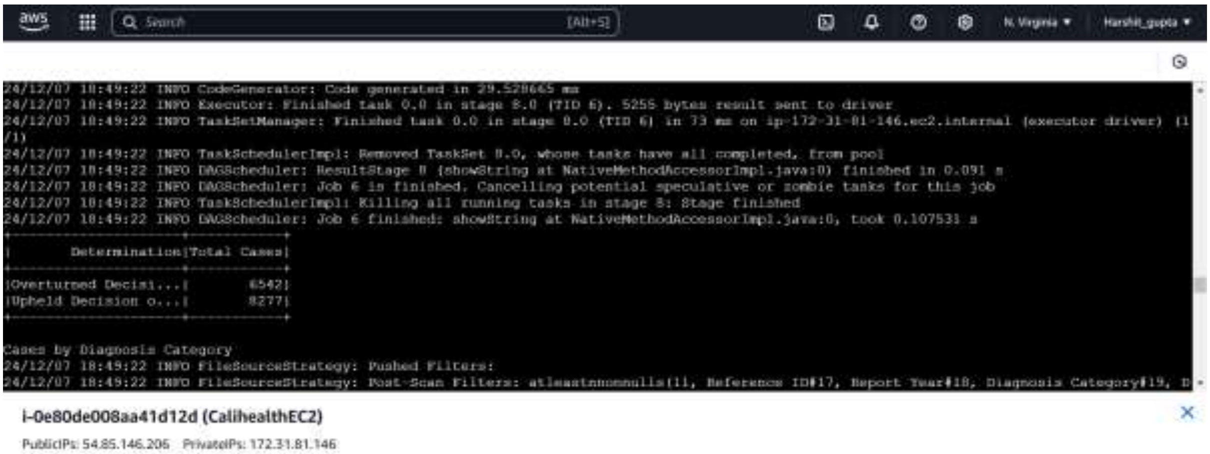


```
24/12/07 18:49:20 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
24/12/07 18:49:20 INFO DAGScheduler: ResultStage 5 (showString at NativeMethodAccessorImpl.java:0) finished in 0.224 s
24/12/07 18:49:20 INFO DAGScheduler: Job 4 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 18:49:20 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
24/12/07 18:49:20 INFO DAGScheduler: Job 4 finished: showString at NativeMethodAccessorImpl.java:0, took 0.259968 s
24/12/07 18:49:20 INFO CodeGenerator: Code generated in 15.96374 ms
+-----------+-----------+
|Report Year|Total Cases|
+-----------+-----------+
|       2003|        710|
|       2007|       1071|
|       2015|       1940|
|       2006|        811|
|       2013|       1087|
|       2014|       1342|
|       2004|        684|
|       2012|       1046|
|       2009|       1031|
|       2016|       1165|
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.85.146.206 PrivateIPs: 172.31.81.146

Distribution of Determination

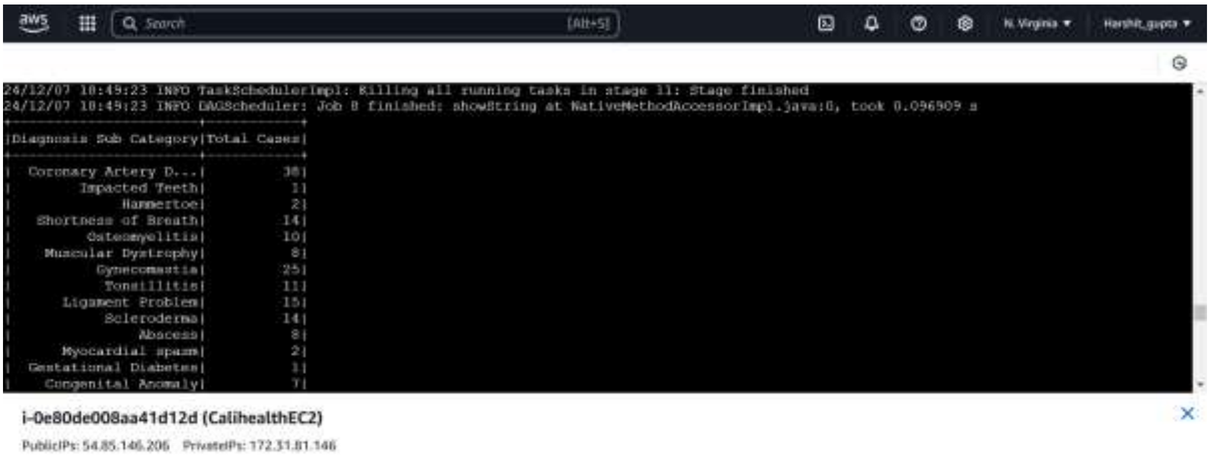


```
24/12/07 18:49:22 INFO CodeGenerator: Code generated in 29.528665 ms
24/12/07 18:49:22 INFO Executor: Finished task 0.0 in stage 8.0 (TID 6). 5255 bytes result sent to driver
24/12/07 18:49:22 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 6) in 73 ms on ip-172-31-81-146.ec2.internal (executor driver) (1
/1)
24/12/07 18:49:22 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
24/12/07 18:49:22 INFO DAGScheduler: ResultStage 8 (showString at NativeMethodAccessorImpl.java:0) finished in 0.091 s
24/12/07 18:49:22 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 18:49:22 INFO TaskSchedulerImpl: Killing all running tasks in stage 8: Stage finished
24/12/07 18:49:22 INFO DAGScheduler: Job 6 finished: showString at NativeMethodAccessorImpl.java:0, took 0.107531 s
+--------------------+-----------+
|       Determination|Total Cases|
+--------------------+-----------+
|Overturned Decisi...|       6542|
|Upheld Decision o...|       8277|
+--------------------+-----------+

Cases by Diagnosis Category
24/12/07 18:49:22 INFO FileSourceStrategy: Pushed Filters:
24/12/07 18:49:22 INFO FileSourceStrategy: Post-Scan Filters: atleastnnonnulls(11, Reference ID#17, Report Year#18, Diagnosis Category#19, D
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.85.146.206 PrivateIPs: 172.31.81.146

Cases by Diagnosis Category

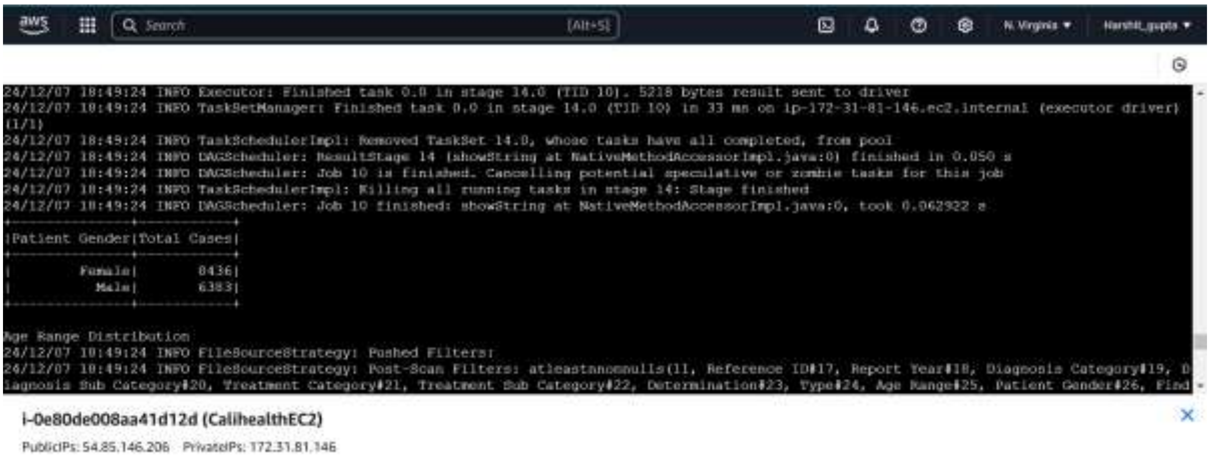


```
24/12/07 18:49:23 INFO TaskSchedulerImpl: Killing all running tasks in stage 11: Stage finished
24/12/07 18:49:23 INFO DAGScheduler: Job 8 finished: showString at NativeMethodAccessorImpl.java:0, took 0.096909 s
+----------------------+-----------+
|Diagnosis Sub Category|Total Cases|
+----------------------+-----------+
|  Coronary Artery D...|         38|
|        Impacted Teeth|          1|
|             Hammertoe|          2|
|   Shortness of Breath|         14|
|          Osteomyelitis|        10|
|     Muscular Dystrophy|         8|
|           Gynecomastia|        25|
|            Tonsillitis|        11|
|       Ligament Problem|        15|
|            Scleroderma|        14|
|                Abscess|         8|
|       Myocardial spasm|         2|
|  Gestational Diabetes|          1|
|    Congenital Anomaly|          7|
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.85.146.206 PrivateIPs: 172.31.81.146

Gender-based distribution of cases

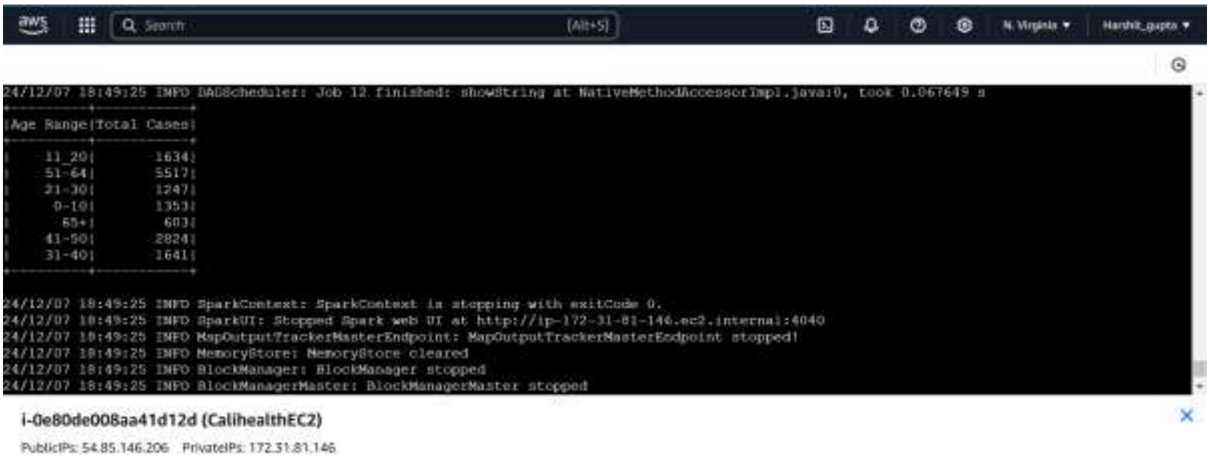


```
24/12/07 18:49:24 INFO Executor: Finished task 0.0 in stage 14.0 (TID 10). 5218 bytes result sent to driver
24/12/07 18:49:24 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 10) in 33 ms on ip-172-31-81-146.ec2.internal (executor driver)
(1/1)
24/12/07 18:49:24 INFO TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
24/12/07 18:49:24 INFO DAGScheduler: ResultStage 14 (showString at NativeMethodAccessorImpl.java:0) finished in 0.050 s
24/12/07 18:49:24 INFO DAGScheduler: Job 10 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 18:49:24 INFO TaskSchedulerImpl: Killing all running tasks in stage 14: Stage finished
24/12/07 18:49:24 INFO DAGScheduler: Job 10 finished: showString at NativeMethodAccessorImpl.java:0, took 0.062922 s
+--------------+-----------+
|Patient Gender|Total Cases|
+--------------+-----------+
|        Female|       8436|
|          Male|       6383|
+--------------+-----------+

Age Range Distribution
24/12/07 18:49:24 INFO FileSourceStrategy: Pushed Filters:
24/12/07 18:49:24 INFO FileSourceStrategy: Post-Scan Filters: atleastnnonnulls(11, Reference ID#17, Report Year#18, Diagnosis Category#19, D
iagnosis Sub Category#20, Treatment Category#21, Treatment Sub Category#22, Determination#23, Type#24, Age Range#25, Patient Gender#26, Find
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.85.146.206 PrivateIPs: 172.31.81.146

Age-range analysis

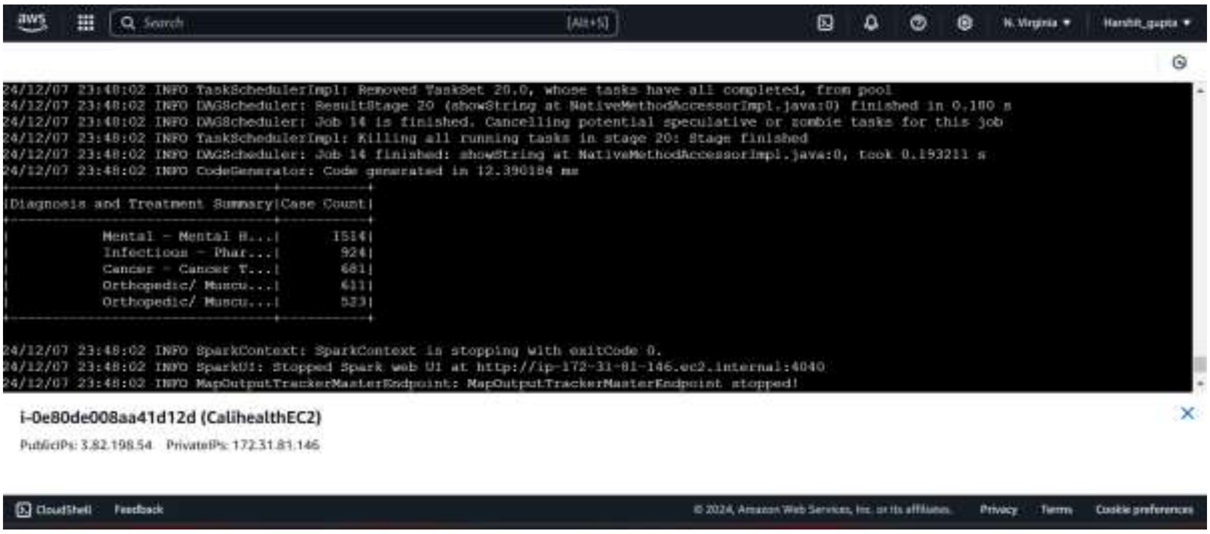


```
24/12/07 18:49:25 INFO DAGScheduler: Job 12 finished: showString at NativeMethodAccessorImpl.java:0, took 0.067649 s
+---------+-----------+
|Age Range|Total Cases|
+---------+-----------+
|    11_20|       1634|
|    51-64|       5517|
|    21-30|       1247|
|     0-10|       1353|
|      65+|        603|
|    41-50|       2824|
|    31-40|       1641|
+---------+-----------+

24/12/07 18:49:25 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/07 18:49:25 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-81-146.ec2.internal:4040
24/12/07 18:49:25 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/07 18:49:25 INFO MemoryStore: MemoryStore cleared
24/12/07 18:49:25 INFO BlockManager: BlockManager stopped
24/12/07 18:49:25 INFO BlockManagerMaster: BlockManagerMaster stopped
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 54.85.146.206 PrivateIPs: 172.31.81.146

Diagnosis and Treatment Summary



```
24/12/07 23:48:02 INFO TaskSchedulerImpl: Removed TaskSet 20.0, whose tasks have all completed, from pool
24/12/07 23:48:02 INFO DAGScheduler: ResultStage 20 (showString at NativeMethodAccessorImpl.java:0) finished in 0.180 s
24/12/07 23:48:02 INFO DAGScheduler: Job 14 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 23:48:02 INFO TaskSchedulerImpl: Killing all running tasks in stage 20: Stage finished
24/12/07 23:48:02 INFO DAGScheduler: Job 14 finished: showString at NativeMethodAccessorImpl.java:0, took 0.193211 s
24/12/07 23:48:02 INFO CodeGenerator: Code generated in 12.390184 ms
+-------------------------------+----------+
|Diagnosis and Treatment Summary|Case Count|
+-------------------------------+----------+
|            Mental - Mental H...|      1514|
|            Infectious - Phar...|       924|
|            Cancer - Cancer T...|       681|
|            Orthopedic/ Muscu...|       611|
|            Orthopedic/ Muscu...|       523|
+-------------------------------+----------+

24/12/07 23:48:02 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/07 23:48:02 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-81-146.ec2.internal:4040
24/12/07 23:48:02 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

i-0e80de008aa41d12d (CalihealthEC2)

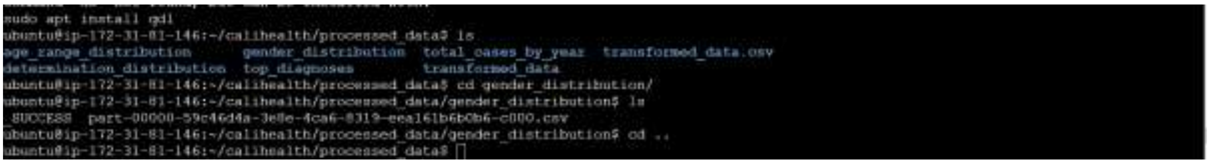
PublicIPs: 3.82.198.54   PrivateIPs: 172.31.81.146

Explanation: Each image captures a different stage of processing—cleaning, transforming, or aggregating—depicting how raw data is converted into a structured and meaningful format suitable for analysis.

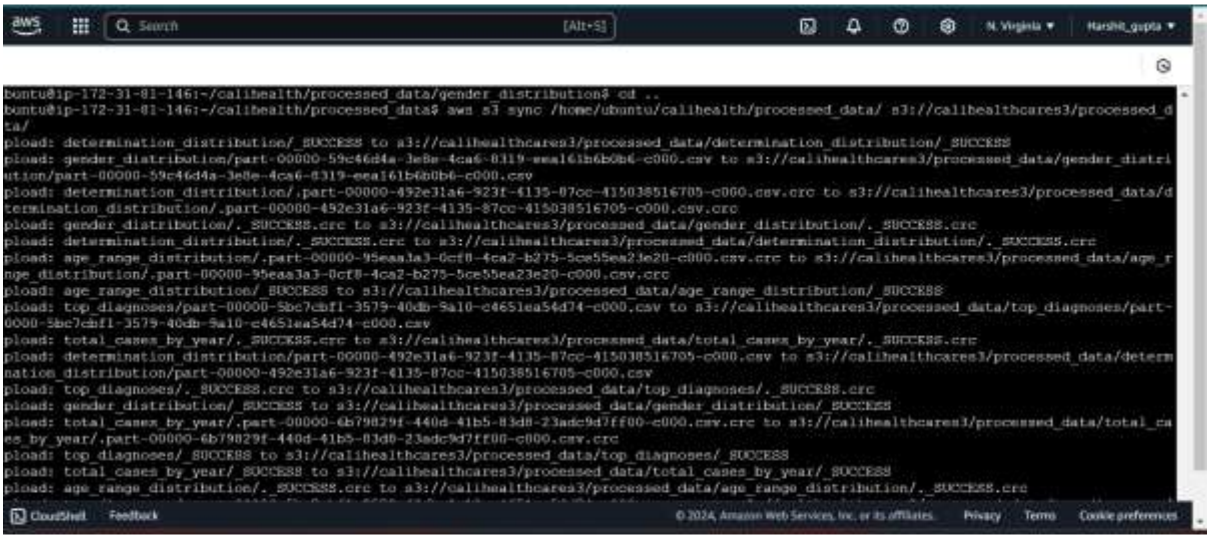
Task 3: Store Processed Data Back to S3 (0.5 Point)

- **Objective:** Save the processed and aggregated data to a new S3 bucket or folder.

- **Steps:** 1. Export data in CSV or Parquet format.



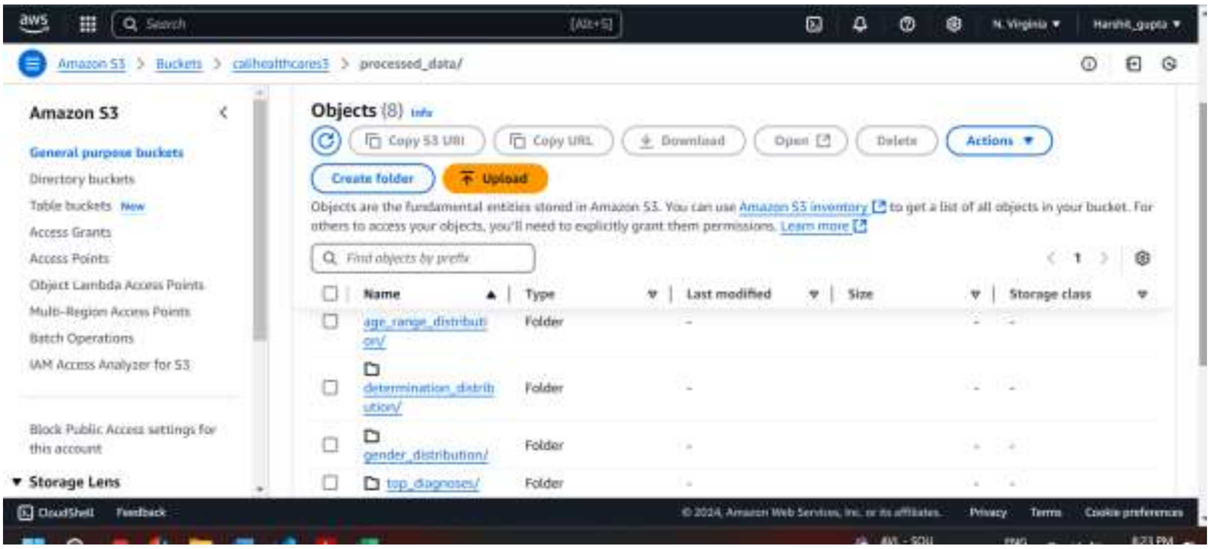
```
sudo apt install gdl
ubuntu@ip-172-31-81-146:~/calihealth/processed_data$ ls
age_range_distribution      gender_distribution  total_cases_by_year  transformed_data.csv
determination_distribution  top_diagnoses        transformed_data
ubuntu@ip-172-31-81-146:~/calihealth/processed_data$ cd gender_distribution/
ubuntu@ip-172-31-81-146:~/calihealth/processed_data/gender_distribution$ ls
_SUCCESS  part-00000-59c46d4a-3e8e-4ca6-8319-eea161b6b0b6-c000.csv
ubuntu@ip-172-31-81-146:~/calihealth/processed_data/gender_distribution$ cd ..
ubuntu@ip-172-31-81-146:~/calihealth/processed_data$ 
```



```
buntu@ip-172-31-81-146:~/calihealth/processed_data/gender_distribution$ cd ..
buntu@ip-172-31-81-146:~/calihealth/processed_data$ aws s3 sync /home/ubuntu/calihealth/processed_data/ s3://calihealthcares3/processed_d
ta/
pload: determination_distribution/_SUCCESS to s3://calihealthcares3/processed_data/determination_distribution/_SUCCESS
pload: gender_distribution/part-00000-59c46d4a-3e8e-4ca6-8319-eea161b6b0b6-c000.csv to s3://calihealthcares3/processed_data/gender_distri
ution/part-00000-59c46d4a-3e8e-4ca6-8319-eea161b6b0b6-c000.csv
pload: determination_distribution/.part-00000-492e31a6-923f-4135-87cc-415038516705-c000.csv.crc to s3://calihealthcares3/processed_data/d
termination_distribution/.part-00000-492e31a6-923f-4135-87cc-415038516705-c000.csv.crc
pload: gender_distribution/._SUCCESS.crc to s3://calihealthcares3/processed_data/gender_distribution/._SUCCESS.crc
pload: determination_distribution/._SUCCESS.crc to s3://calihealthcares3/processed_data/determination_distribution/._SUCCESS.crc
pload: age_range_distribution/.part-00000-95eaa3a3-0cf8-4ca2-b275-5ce55ea23e20-c000.csv.crc to s3://calihealthcares3/processed_data/age_r
nge_distribution/.part-00000-95eaa3a3-0cf8-4ca2-b275-5ce55ea23e20-c000.csv.crc
pload: age_range_distribution/_SUCCESS to s3://calihealthcares3/processed_data/age_range_distribution/_SUCCESS
pload: top_diagnoses/part-00000-5bc7cbf1-3579-40db-9a10-c4651ea54d74-c000.csv to s3://calihealthcares3/processed_data/top_diagnoses/part-
0000-5bc7cbf1-3579-40db-9a10-c4651ea54d74-c000.csv
pload: total_cases_by_year/._SUCCESS.crc to s3://calihealthcares3/processed_data/total_cases_by_year/._SUCCESS.crc
pload: determination_distribution/part-00000-492e31a6-923f-4135-87cc-415038516705-c000.csv to s3://calihealthcares3/processed_data/determ
nation_distribution/part-00000-492e31a6-923f-4135-87cc-415038516705-c000.csv
pload: top_diagnoses/._SUCCESS.crc to s3://calihealthcares3/processed_data/top_diagnoses/._SUCCESS.crc
pload: gender_distribution/_SUCCESS to s3://calihealthcares3/processed_data/gender_distribution/_SUCCESS
pload: total_cases_by_year/.part-00000-6b79829f-440d-41b5-83d8-23adc9d7ff00-c000.csv.crc to s3://calihealthcares3/processed_data/total_ca
es_by_year/.part-00000-6b79829f-440d-41b5-83d8-23adc9d7ff00-c000.csv.crc
pload: top_diagnoses/_SUCCESS to s3://calihealthcares3/processed_data/top_diagnoses/_SUCCESS
pload: total_cases_by_year/_SUCCESS to s3://calihealthcares3/processed_data/total_cases_by_year/_SUCCESS
pload: age_range_distribution/._SUCCESS.crc to s3://calihealthcares3/processed_data/age_range_distribution/._SUCCESS.crc
```

The image should illustrate the use of PySpark's write.csv or write.parquet commands and subsequent confirmation of the file upload in the S3 bucket. This ensures the processed data is stored securely and is accessible for later analysis.

2. Upload the processed data to a designated S3 location for easy access. (Generating new CSV, downloading it and uploading to S3 bucket is fine).

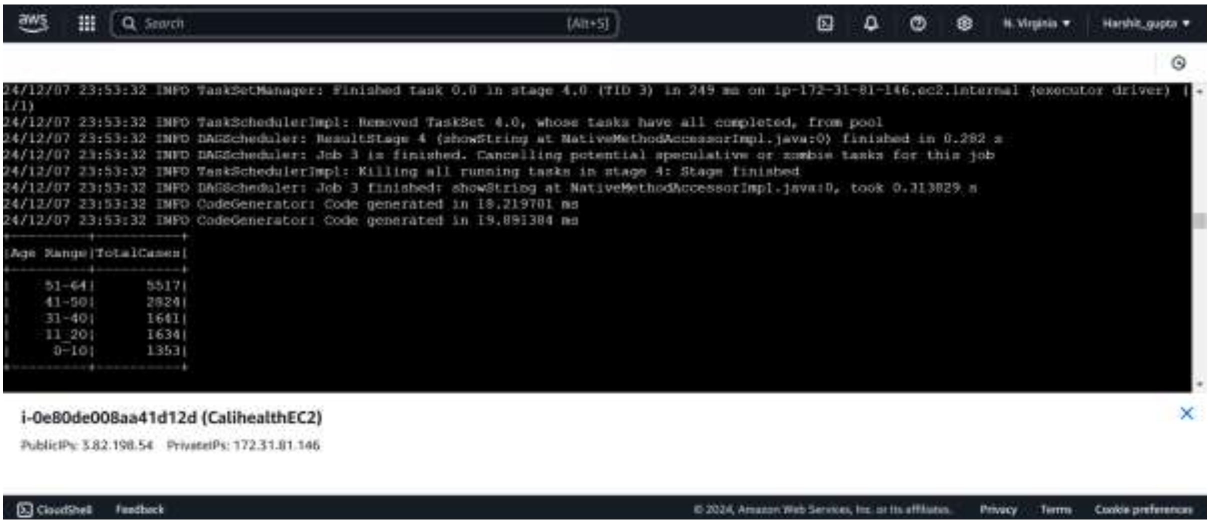


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	age_range_distribution/	Folder	-	-	-
□	determination_distribution/	Folder	-	-	-
□	gender_distribution/	Folder	-	-	-
□	top_diagnoses/	Folder	-	-	-

Task 4: Data Analysis Using Spark SQL (1 Point)

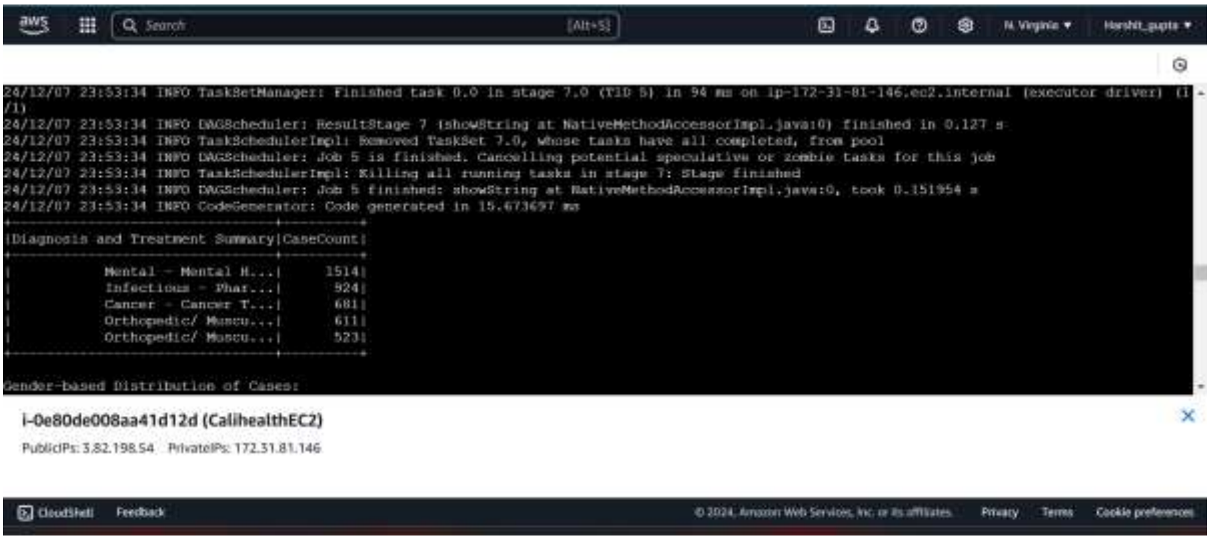
- **Objective:** Use SQL to derive insights (Atleast 5 Queries).
- **Example Queries:** 1. Identify top-performing regions. 2. Analyze month-over-month revenue growth. 3. Determine the most popular product categories.



```
24/12/07 23:53:32 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 3) in 249 ms on ip-172-31-81-146.ec2.internal (executor driver) (
1/1)
24/12/07 23:53:32 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool
24/12/07 23:53:32 INFO DAGScheduler: ResultStage 4 (showString at NativeMethodAccessorImpl.java:0) finished in 0.282 s
24/12/07 23:53:32 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 23:53:32 INFO TaskSchedulerImpl: Killing all running tasks in stage 4: Stage finished
24/12/07 23:53:32 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.313829 s
24/12/07 23:53:32 INFO CodeGenerator: Code generated in 18.219701 ms
24/12/07 23:53:32 INFO CodeGenerator: Code generated in 19.891384 ms
+---------+----------+
|Age Range|TotalCases|
+---------+----------+
|    51-64|      5517|
|    41-50|      2824|
|    31-40|      1641|
|    11_20|      1634|
|     0-10|      1353|
+---------+----------+
```

i-0e80de008aa41d12d (CalihealthEC2)

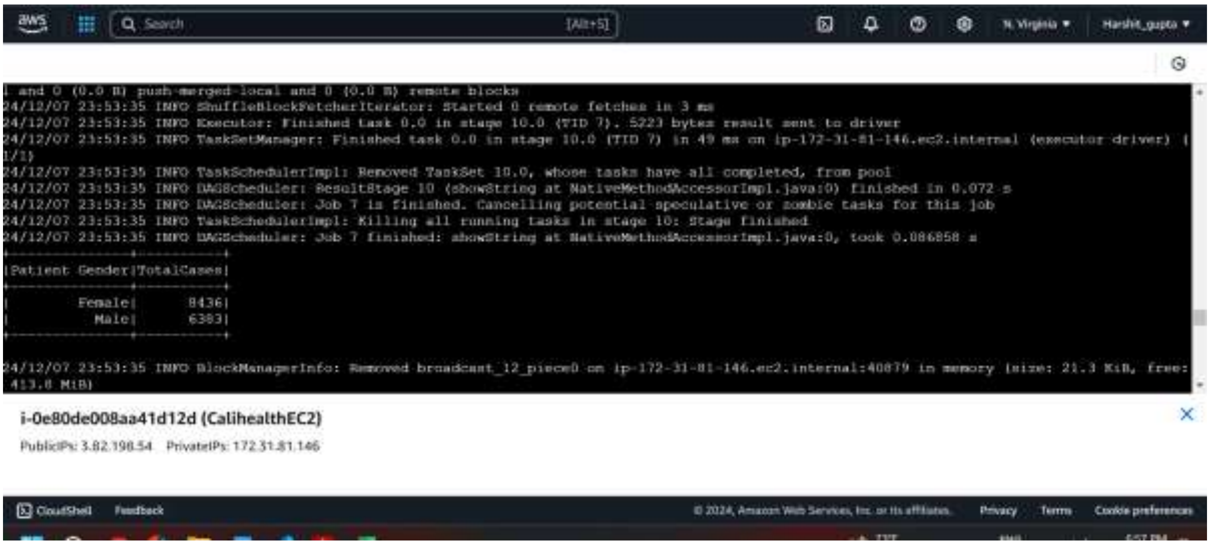
PublicIPs: 3.82.198.54   PrivateIPs: 172.31.81.146



```
24/12/07 23:53:34 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 5) in 94 ms on ip-172-31-81-146.ec2.internal (executor driver) (1
/1)
24/12/07 23:53:34 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0) finished in 0.127 s
24/12/07 23:53:34 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
24/12/07 23:53:34 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 23:53:34 INFO TaskSchedulerImpl: Killing all running tasks in stage 7: Stage finished
24/12/07 23:53:34 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.151954 s
24/12/07 23:53:34 INFO CodeGenerator: Code generated in 15.673697 ms
+-------------------------------+---------+
|Diagnosis and Treatment Summary|CaseCount|
+-------------------------------+---------+
|           Mental - Mental H...|     1514|
|           Infectious - Phar...|      924|
|           Cancer - Cancer T...|      681|
|           Orthopedic/ Muscu...|      611|
|           Orthopedic/ Muscu...|      523|
+-------------------------------+---------+

Gender-based Distribution of Cases:
```

i-0e80de008aa41d12d (CalihealthEC2)

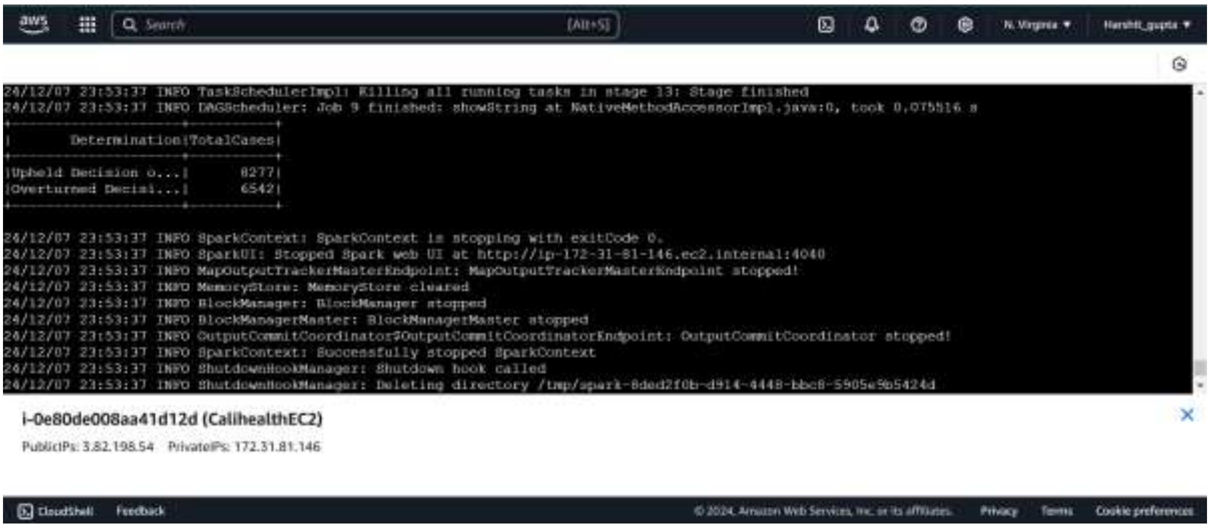
PublicIPs: 3.82.198.54   PrivateIPs: 172.31.81.146



```
1 and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
24/12/07 23:53:35 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
24/12/07 23:53:35 INFO Executor: Finished task 0.0 in stage 10.0 (TID 7). 5223 bytes result sent to driver
24/12/07 23:53:35 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 7) in 49 ms on ip-172-31-81-146.ec2.internal (executor driver) (
1/1)
24/12/07 23:53:35 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
24/12/07 23:53:35 INFO DAGScheduler: ResultStage 10 (showString at NativeMethodAccessorImpl.java:0) finished in 0.072 s
24/12/07 23:53:35 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/07 23:53:35 INFO TaskSchedulerImpl: Killing all running tasks in stage 10: Stage finished
24/12/07 23:53:35 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.086858 s
+--------------+----------+
|Patient Gender|TotalCases|
+--------------+----------+
|        Female|      8436|
|          Male|      6383|
+--------------+----------+

24/12/07 23:53:35 INFO BlockManagerInfo: Removed broadcast_12_piece0 on ip-172-31-81-146.ec2.internal:40879 in memory (size: 21.3 KiB, free:
 413.8 MiB)
```

i-0e80de008aa41d12d (CalihealthEC2)

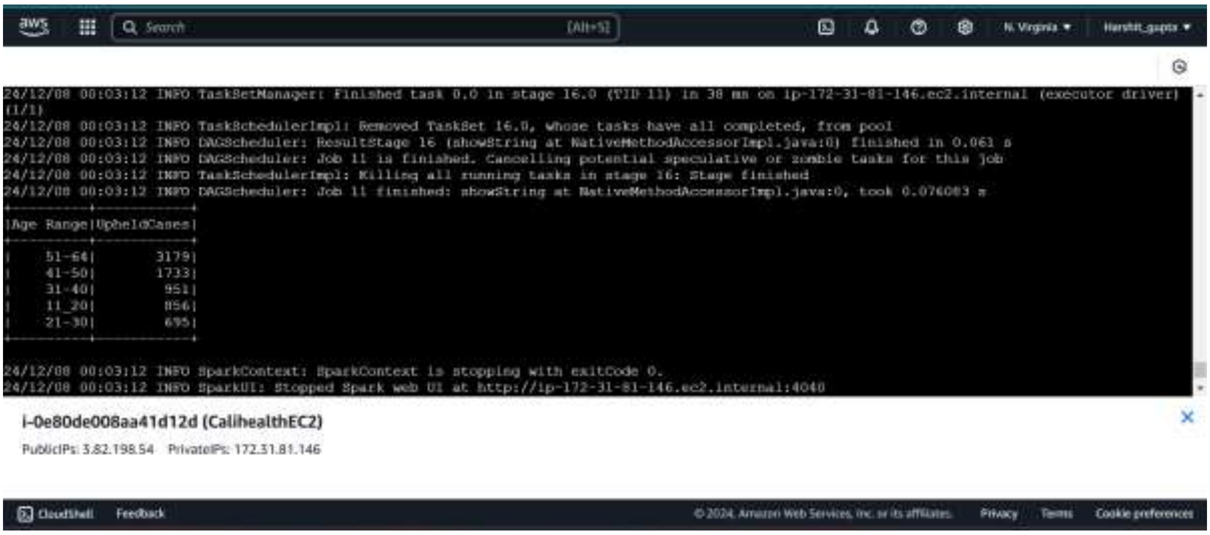
PublicIPs: 3.82.198.54   PrivateIPs: 172.31.81.146



```
24/12/07 23:53:37 INFO TaskSchedulerImpl: Killing all running tasks in stage 13: Stage finished
24/12/07 23:53:37 INFO DAGScheduler: Job 9 finished: showString at NativeMethodAccessorImpl.java:0, took 0.075516 s
+--------------------+----------+
|       Determination|TotalCases|
+--------------------+----------+
|Upheld Decision o...|      8277|
|Overturned Decisi...|      6542|
+--------------------+----------+

24/12/07 23:53:37 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/07 23:53:37 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-81-146.ec2.internal:4040
24/12/07 23:53:37 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/07 23:53:37 INFO MemoryStore: MemoryStore cleared
24/12/07 23:53:37 INFO BlockManager: BlockManager stopped
24/12/07 23:53:37 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/07 23:53:37 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/07 23:53:37 INFO SparkContext: Successfully stopped SparkContext
24/12/07 23:53:37 INFO ShutdownHookManager: Shutdown hook called
24/12/07 23:53:37 INFO ShutdownHookManager: Deleting directory /tmp/spark-8ded2f0b-d914-4448-bbc8-5905e9b5424d
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 3.82.198.54 PrivateIPs: 172.31.81.146



```
24/12/08 00:03:12 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 11) in 38 ms on ip-172-31-81-146.ec2.internal (executor driver)
(1/1)
24/12/08 00:03:12 INFO TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
24/12/08 00:03:12 INFO DAGScheduler: ResultStage 16 (showString at NativeMethodAccessorImpl.java:0) finished in 0.061 s
24/12/08 00:03:12 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/08 00:03:12 INFO TaskSchedulerImpl: Killing all running tasks in stage 16: Stage finished
24/12/08 00:03:12 INFO DAGScheduler: Job 11 finished: showString at NativeMethodAccessorImpl.java:0, took 0.076083 s
+---------+-----------+
|Age Range|UpheldCases|
+---------+-----------+
|    51-64|       3179|
|    41-50|       1733|
|    31-40|        951|
|    11_20|        856|
|    21-30|        695|
+---------+-----------+

24/12/08 00:03:12 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/08 00:03:12 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-81-146.ec2.internal:4040
```

i-0e80de008aa41d12d (CalihealthEC2)

PublicIPs: 3.82.198.54 PrivateIPs: 172.31.81.146

Each image depicts a complete SQL workflow—from formulating queries to interpreting results. This showcases how Spark SQL facilitates powerful, SQL-based data exploration.

Task 5: Machine Learning with AWS SageMaker Autopilot (1.5 Point)

● **Objective:** Use **AWS SageMaker Autopilot** to train and evaluate machine learning models on the **processed dataset** stored in S3 without writing any code.

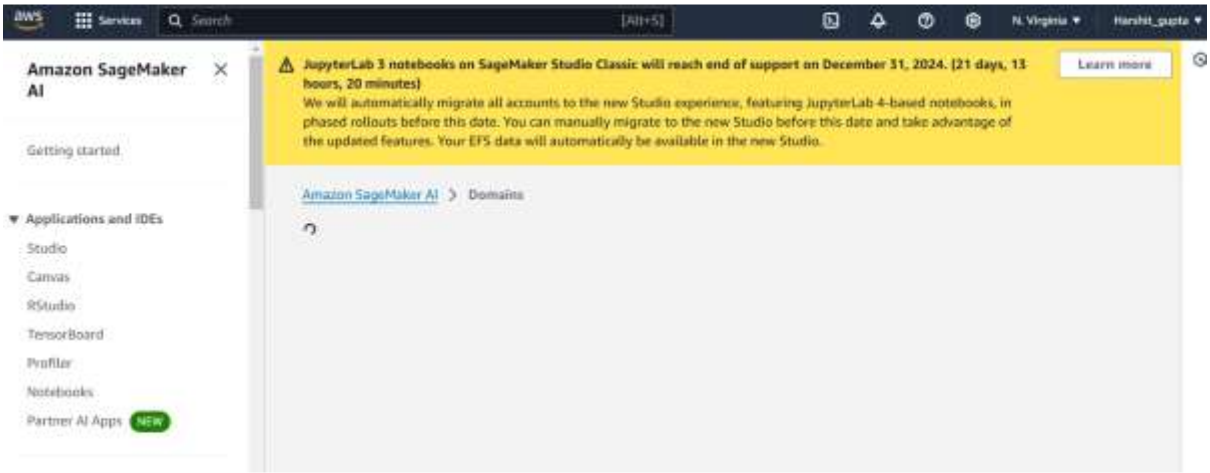
● Here is a tutorial on how to use AWS SageMaker AutoPilot

● **Steps:** 1. **Import Processed Data** : Load the processed dataset from S3 into SageMaker Autopilot.

2. **Run Autopilot Experiment** : ■ Select the target column. ■ Run the AutoML process to train and evaluate multiple models.

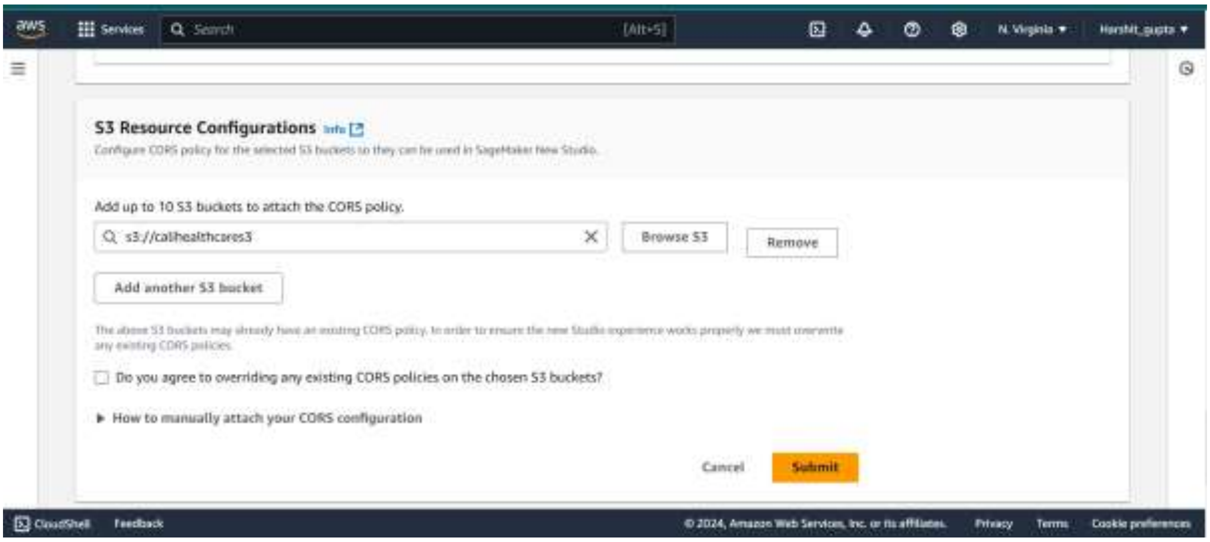
3. Consider taking screenshots of your working and query results.

4. **Review Results** : Analyze the model leaderboard and performance metrics. Address ethical issues like bias in training data and privacy concerns.



JupyterLab 3 notebooks on SageMaker Studio Classic will reach end of support on December 31, 2024. (21 days, 13 hours, 20 minutes)

We will automatically migrate all accounts to the new Studio experience, featuring JupyterLab 4-based notebooks, in phased rollouts before this date. You can manually migrate to the new Studio before this date and take advantage of the updated features. Your EFS data will automatically be available in the new Studio.



S3 Resource Configurations

Configure CORS policy for the selected S3 buckets so they can be used in SageMaker New Studio.

Add up to 10 S3 buckets to attach the CORS policy.

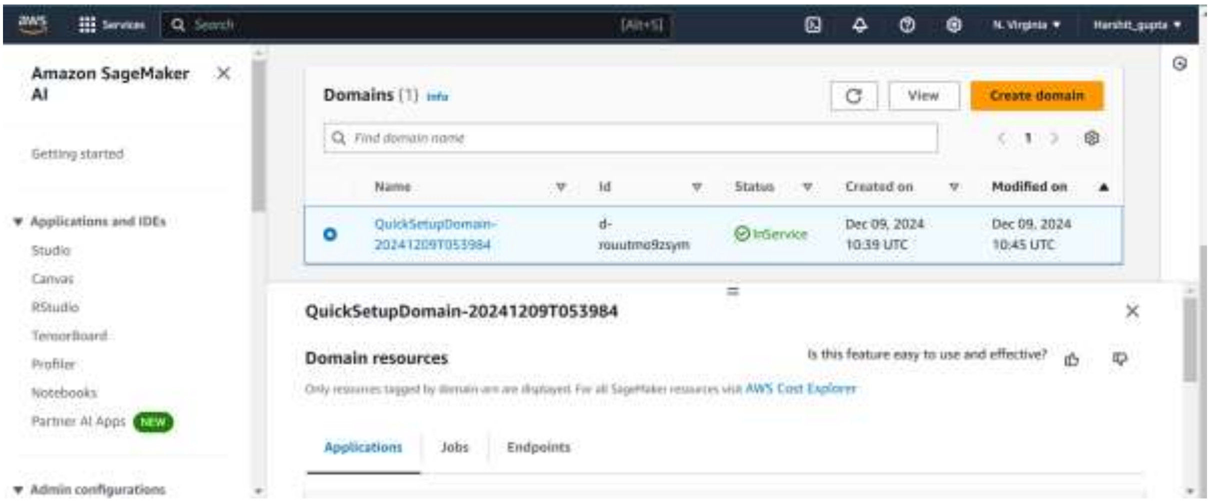
s3://calhealthcares3

The above S3 buckets may already have an existing CORS policy. In order to ensure the new Studio experience works properly we must overwrite any existing CORS policies.

□ Do you agree to overriding any existing CORS policies on the chosen S3 buckets?

▸ How to manually attach your CORS configuration

© 2024, Amazon Web Services, Inc. or its affiliates.



Amazon SageMaker AI

Domains (1) Info

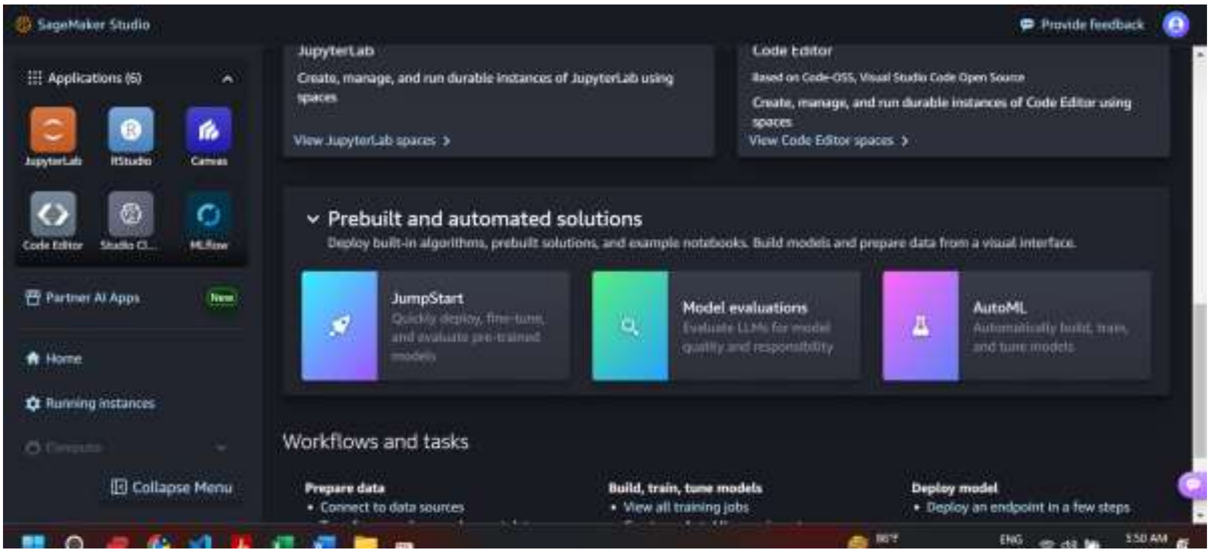
	Name	Id	Status	Created on	Modified on
●	QuickSetupDomain-20241209T053984	d-rouutmo9zsym	✓ InService	Dec 09, 2024 10:39 UTC	Dec 09, 2024 10:45 UTC

QuickSetupDomain-20241209T053984

Domain resources

Is this feature easy to use and effective?

Only resources tagged by domain-arn are displayed. For all SageMaker resources visit AWS Cost Explorer



JupyterLab

Create, manage, and run durable instances of JupyterLab using spaces

View JupyterLab spaces >

Code Editor

Based on Code-OSS, Visual Studio Code Open Source

Create, manage, and run durable instances of Code Editor using spaces

View Code Editor spaces >

Prebuilt and automated solutions

Deploy built-in algorithms, prebuilt solutions, and example notebooks. Build models and prepare data from a visual interface.

JumpStart
Quickly deploy, fine-tune, and evaluate pre-trained models

Model evaluations
Evaluate LLMs for model quality and responsibility

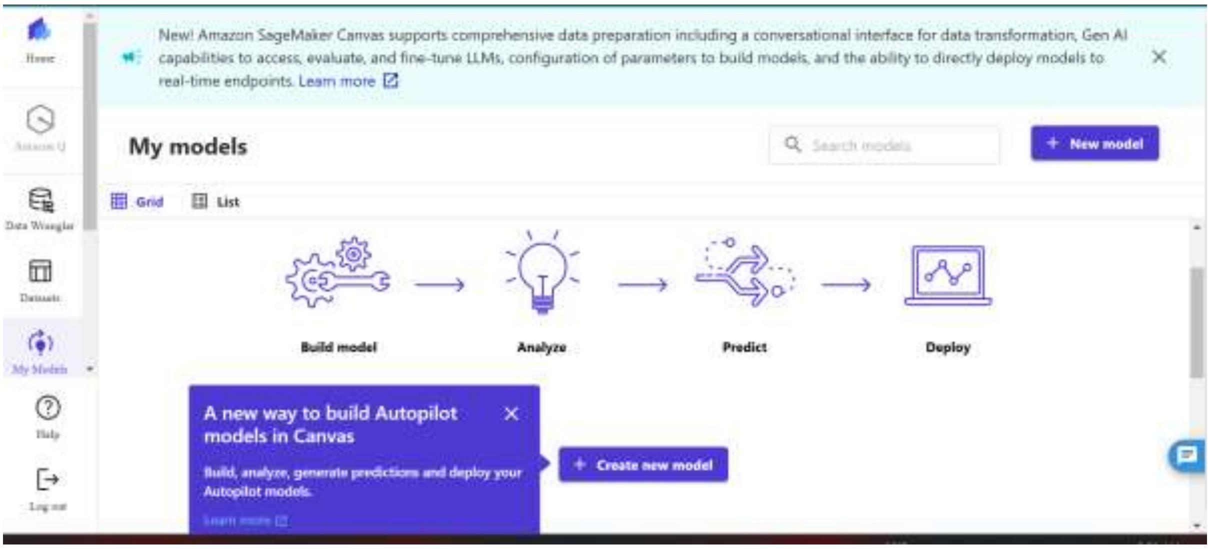
AutoML
Automatically build, train, and tune models

Workflows and tasks

Prepare data
- Connect to data sources

Build, train, tune models
- View all training jobs

Deploy model
- Deploy an endpoint in a few steps

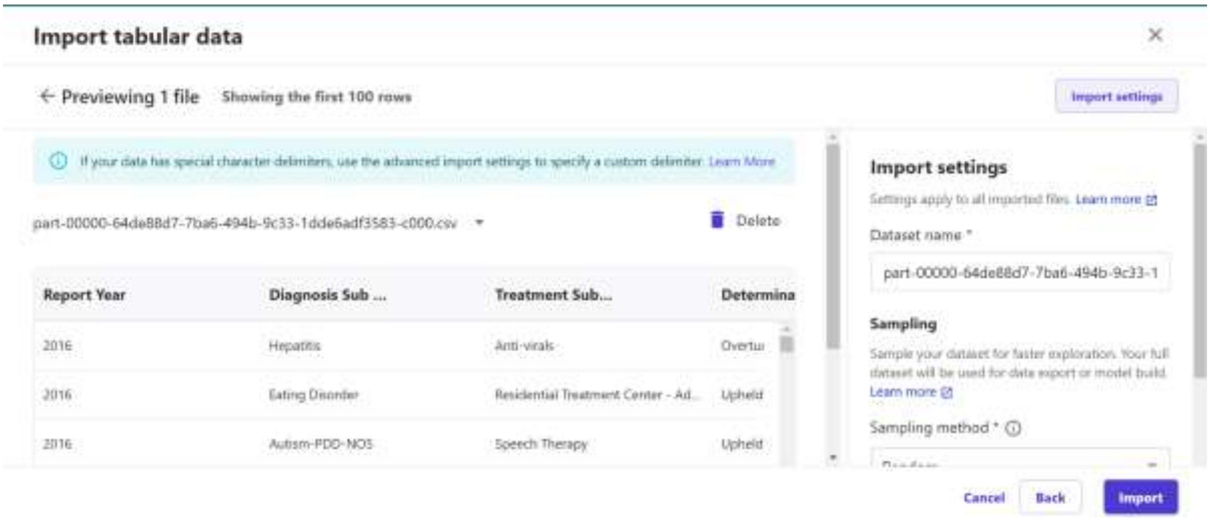


New! Amazon SageMaker Canvas supports comprehensive data preparation including a conversational interface for data transformation, Gen AI capabilities to access, evaluate, and fine-tune LLMs, configuration of parameters to build models, and the ability to directly deploy models to real-time endpoints. Learn more

My models

A new way to build Autopilot models in Canvas

Build, analyze, generate predictions and deploy your Autopilot models.



Import tabular data

← Previewing 1 file Showing the first 100 rows

ⓘ If your data has special character delimiters, use the advanced import settings to specify a custom delimiter. Learn More

part-00000-64de88d7-7ba6-494b-9c33-1dde6adf3583-c000.csv

Report Year	Diagnosis Sub ...	Treatment Sub...	Determina
2016	Hepatitis	Anti-virals	Overtu
2016	Eating Disorder	Residential Treatment Center - Ad...	Upheld
2016	Autism-PDD-NOS	Speech Therapy	Upheld

Import settings

Settings apply to all imported files. Learn more

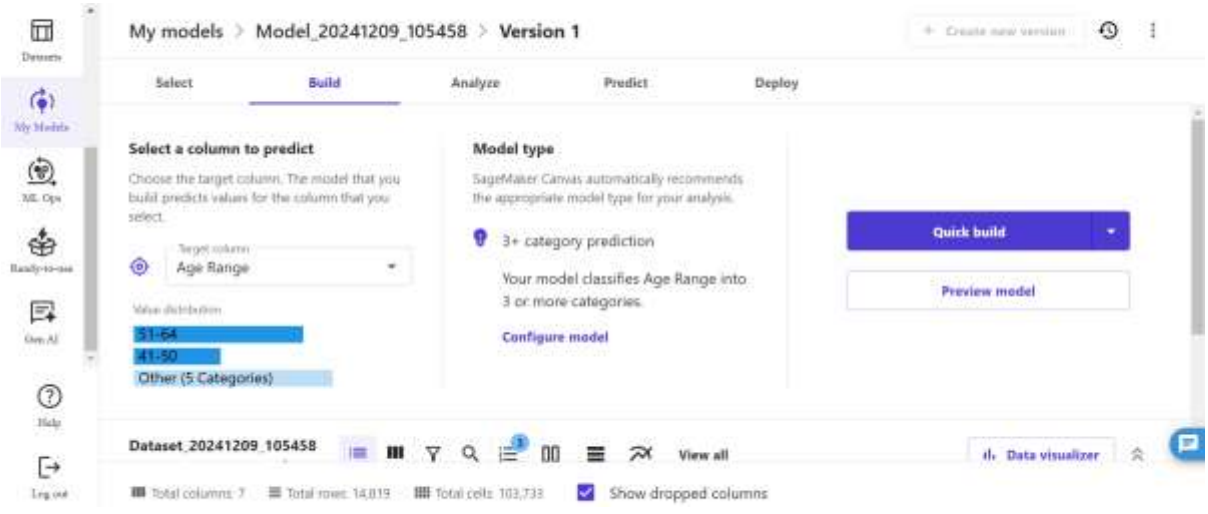
Dataset name *

part-00000-64de88d7-7ba6-494b-9c33-1

Sampling

Sample your dataset for faster exploration. Your full dataset will be used for data export or model build. Learn more

Sampling method * ⓘ



My models > Model_20241209_105458 > Version 1

Select a column to predict

Choose the target column. The model that you build predicts values for the column that you select.

Target column: Age Range

Value distribution

Model type

SageMaker Canvas automatically recommends the appropriate model type for your analysis.

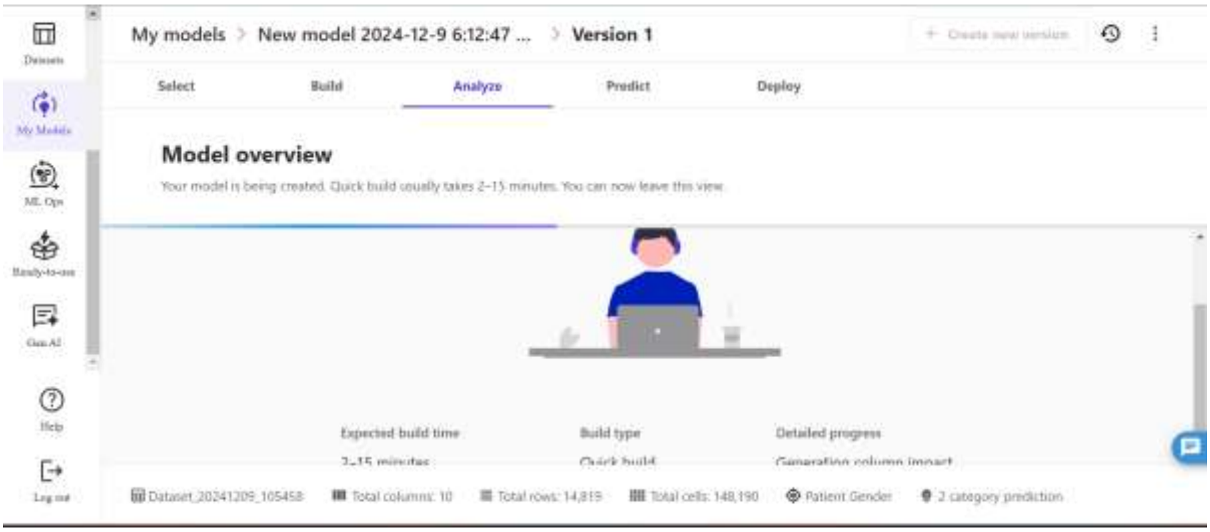
3+ category prediction

Your model classifies Age Range into 3 or more categories.

Configure model

Dataset_20241209_105458

Total columns: 7 Total rows: 14,819 Total cells: 103,733 ☑ Show dropped columns



My models > New model 2024-12-9 6:12:47 ... > Version 1

Model overview

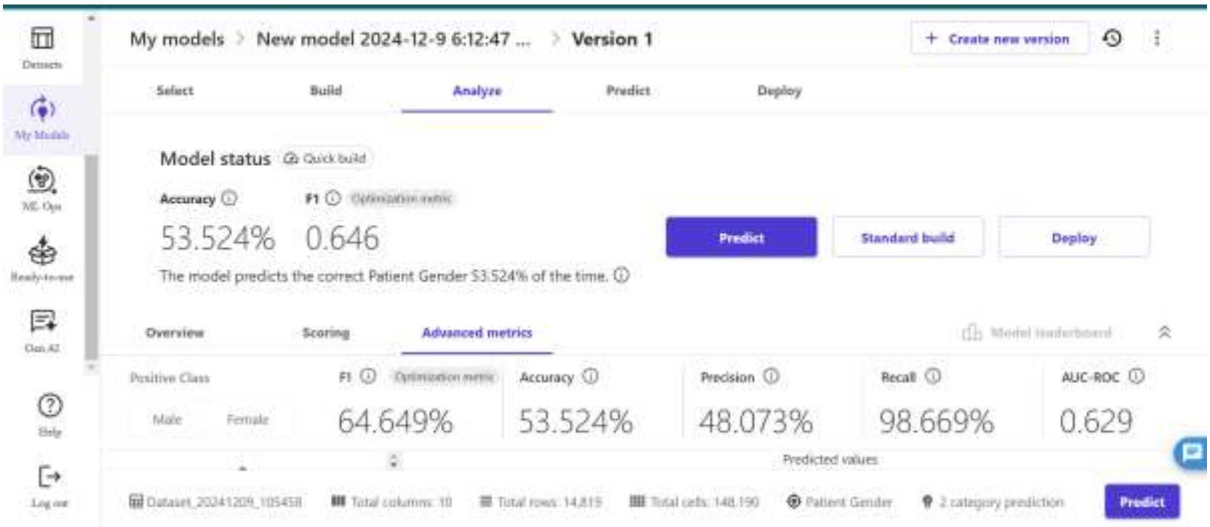
Your model is being created. Quick build usually takes 2–15 minutes. You can now leave this view.

Expected build time: 2–15 minutes

Build type: Quick build

Detailed progress: Generation column impact

Dataset_20241209_105458 Total columns: 10 Total rows: 14,819 Total cells: 148,190 Patient Gender 2 category prediction



Model status Quick build

Accuracy ⓘ 53.524%

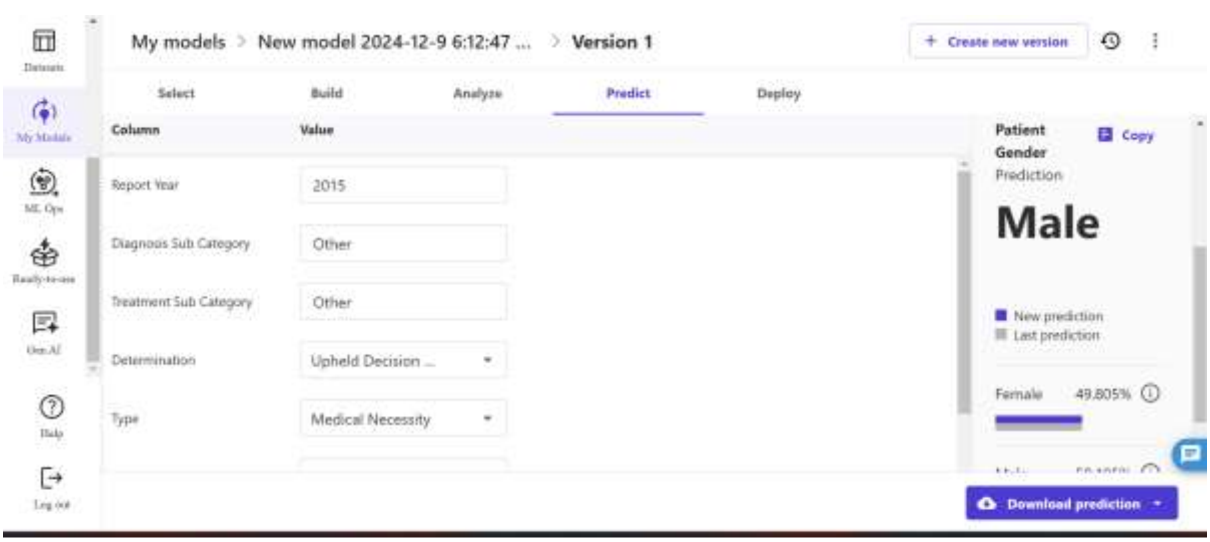
F1 ⓘ Optimization metric 0.646

The model predicts the correct Patient Gender 53.524% of the time. ⓘ

Positive Class	F1 ⓘ Optimization metric	Accuracy ⓘ	Precision ⓘ	Recall ⓘ	AUC-ROC ⓘ
Male Female	64.649%	53.524%	48.073%	98.669%	0.629

Predicted values

Dataset_20241209_105458 | Total columns: 10 | Total rows: 14,819 | Total cells: 148,190 | Patient Gender | 2 category prediction



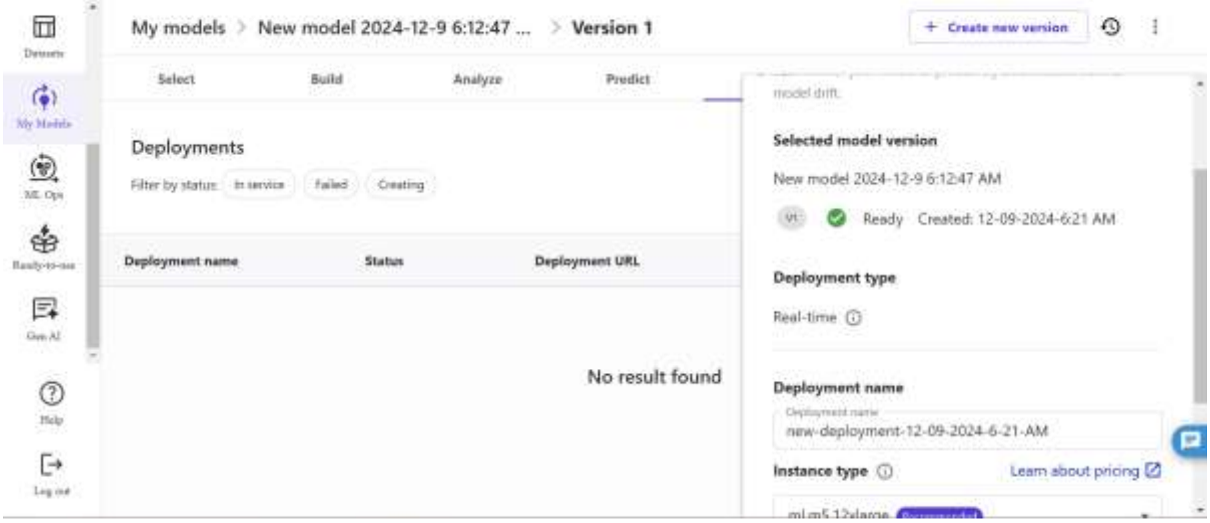
Column	Value
Report Year	2015
Diagnosis Sub Category	Other
Treatment Sub Category	Other
Determination	Upheld Decision ...
Type	Medical Necessity

Patient Gender Prediction

Male

■ New prediction
■ Last prediction

Female 49.805% ⓘ



Deployments

Filter by status: In service Failed Creating

Deployment name	Status	Deployment URL

No result found

model drift.

Selected model version

New model 2024-12-9 6:12:47 AM

V1 ✓ Ready Created: 12-09-2024-6:21 AM

Deployment type

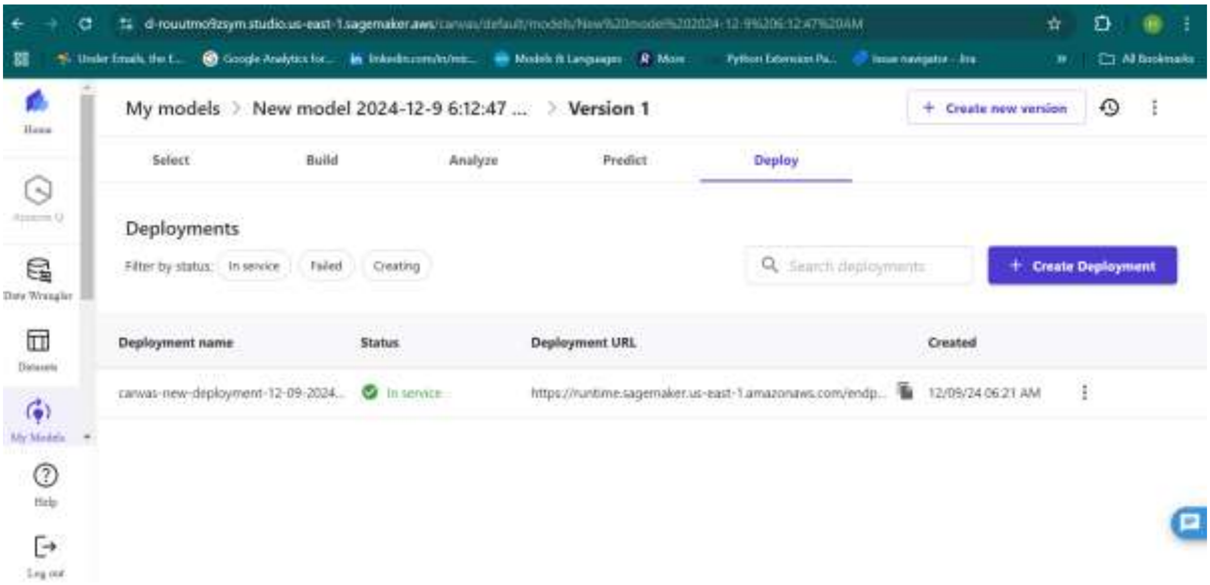
Real-time ⓘ

Deployment name

Deployment name: new-deployment-12-09-2024-6-21-AM

Instance type ⓘ

ml.m5.12xlarge Recommended

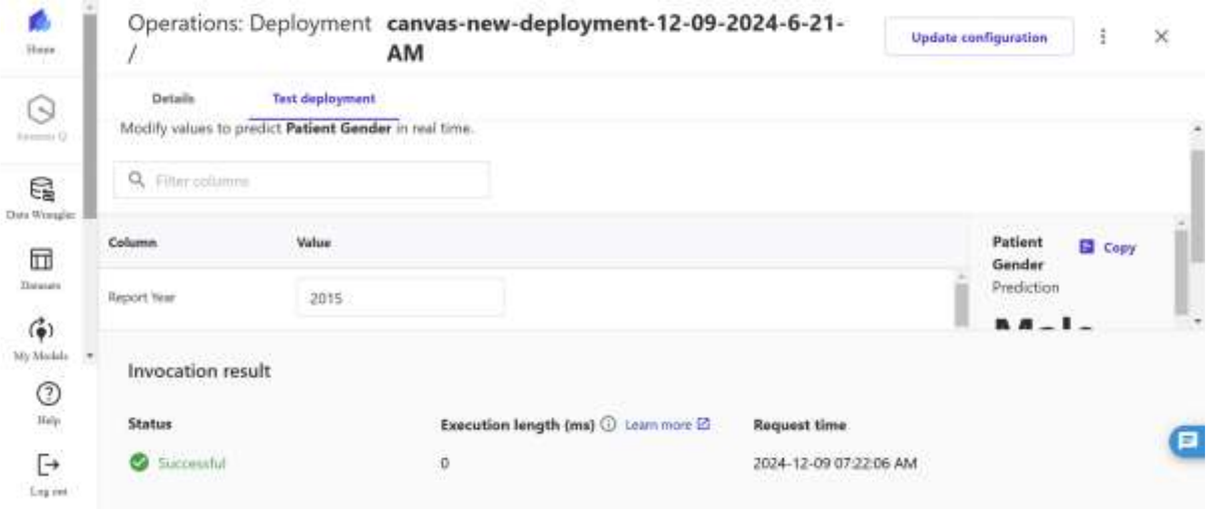


My models > New model 2024-12-9 6:12:47 ... > Version 1

Deployments

Filter by status: In service Failed Creating

Deployment name	Status	Deployment URL	Created
canvas-new-deployment-12-09-2024...	In service	https://runtime.sagemaker.us-east-1.amazonaws.com/endp...	12/09/24 06:21 AM



Operations: Deployment / **canvas-new-deployment-12-09-2024-6-21-AM**

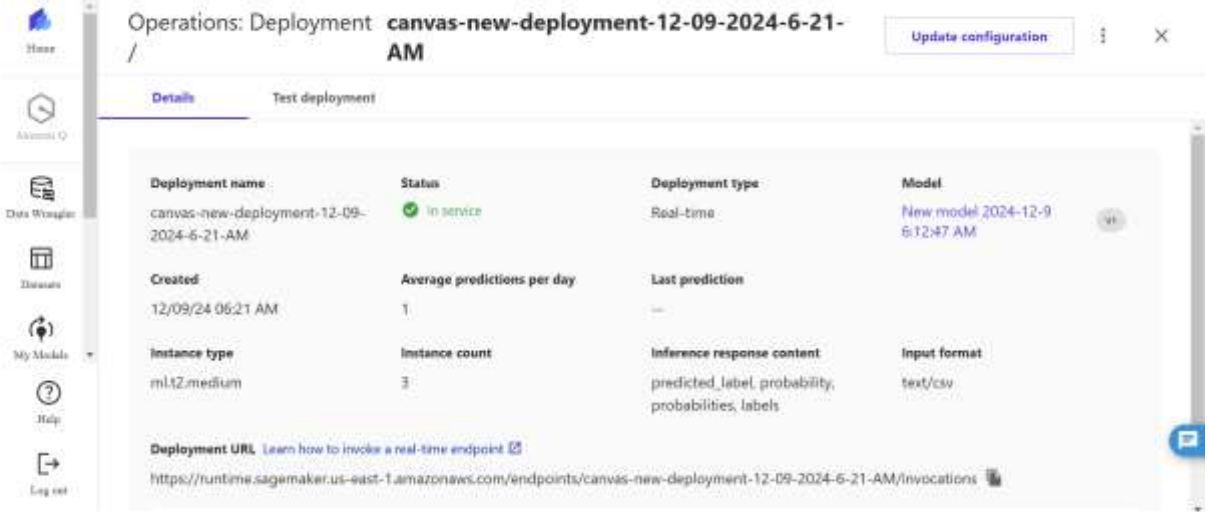
Modify values to predict **Patient Gender** in real time.

Column	Value
Report Year	2015

Patient Gender Prediction

Invocation result

Status	Execution length (ms)	Request time
Successful	0	2024-12-09 07:22:06 AM



Operations: Deployment / **canvas-new-deployment-12-09-2024-6-21-AM**

Deployment name: canvas-new-deployment-12-09-2024-6-21-AM

Status: In service

Deployment type: Real-time

Model: New model 2024-12-9 6:12:47 AM

Created: 12/09/24 06:21 AM

Average predictions per day: 1

Last prediction: –

Instance type: ml.t2.medium

Instance count: 3

Inference response content: predicted_label, probability, probabilities, labels

Input format: text/csv

Deployment URL Learn how to invoke a real-time endpoint

https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/canvas-new-deployment-12-09-2024-6-21-AM/invocations

Image 1: Captures the data import process into SageMaker Autopilot, with the processed dataset loaded from S3.

Image 2: Shows the experiment configuration, including the selected target column and AutoML settings.

Image 3: Highlights the experiment results, such as the model leaderboard and performance metrics.

Explanation: The images provide a step-by-step overview of SageMaker Autopilot, demonstrating its ability to automate model training, evaluation, and comparison, while ensuring no coding is required. Insights into ethical considerations, such as addressing data bias, may also be visualized**.**

4. Visualization (1.5 Point)

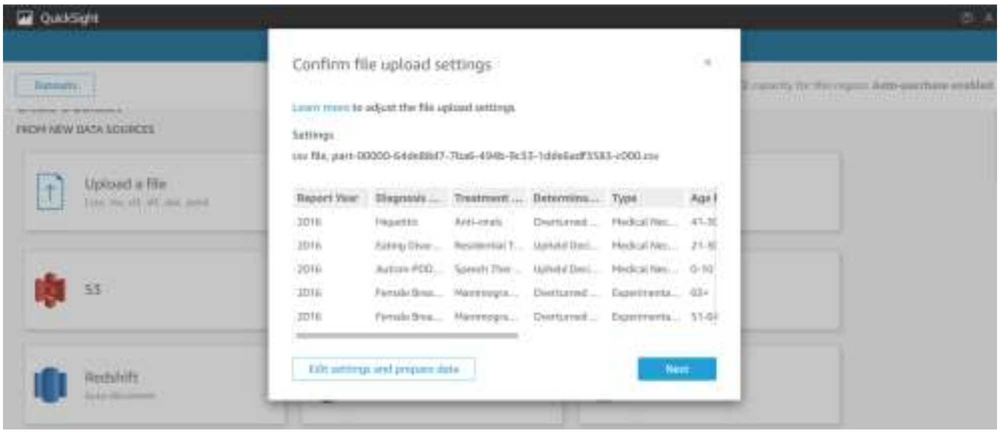
Tool Options:

- **AWS QuickSight** for creating dynamic dashboards. (Note: AWS QuickSight offers a Free Tier for new users. The Free Tier includes the following: 1 GB of SPICE capacity. Here is a tutorial on how to use AWS QuickSight .)
  Task: Create Dashboards

1. Connect QuickSight to the processed data in S3.

https://us-east-1.quicksight.aws.amazon.com/sn/accounts/914497062970/dashboards/d529018e-86d7-477c-8885-06e0a4a57962?directory_alias=hhgupta

2. Design a dashboard with at least 4 insightful visualizations. **Caution:** ■ Be mindful of **QuickSight usage limits** as it may incur additional costs. ■ Use Power BI for local visualization to avoid QuickSight charges.



Confirm file upload settings

Learn more to adjust the file upload settings

Settings

csv file, part-00000-64de88d7-7ba6-494b-9c53-1dde6adf3583-c000.csv

Report Year	Diagnosis ...	Treatment ...	Determina...	Type	Age I
2016	Hepatitis	Anti-virals	Overturned ...	Medical Nec...	41-50
2016	Eating Disor...	Residential T...	Upheld Deci...	Medical Nec...	21-30
2016	Autism-PDD...	Speech Ther...	Upheld Deci...	Medical Nec...	0-10
2016	Female Brea...	Mammogra...	Overturned ...	Experimenta...	65+
2016	Female Brea...	Mammogra...	Overturned ...	Experimenta...	51-64

1. Count of Records by Age Range and Determination

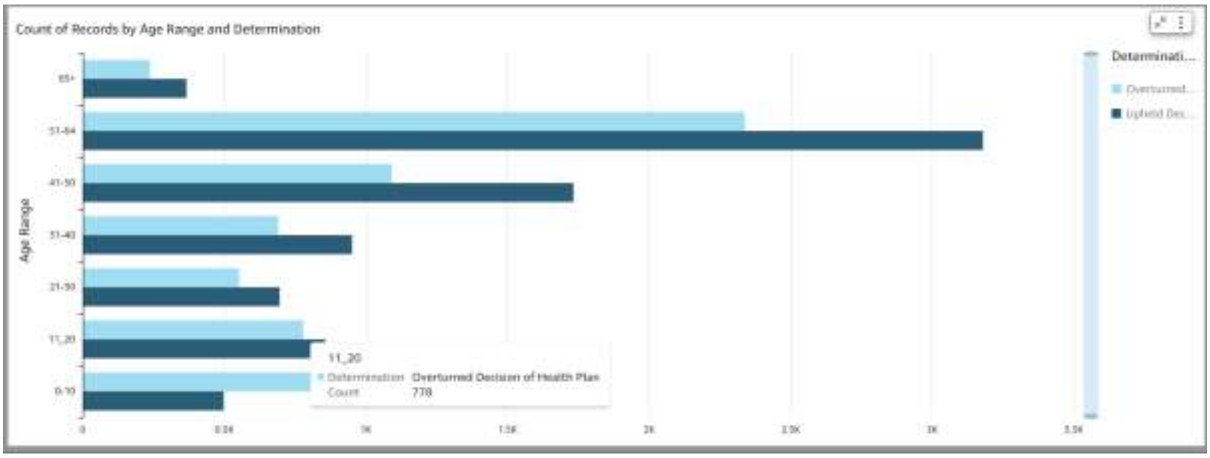
Purpose:
This bar chart visualizes the distribution of records across various age ranges, segmented by the determination outcome (e.g., "Overturned Decision of Health Plan" and "Upheld Decision of Health Plan").

Insights Derived:

- Higher concentrations of records are seen in the age range 51-64, indicating this group undergoes more health plan determinations.
- The "Overturned Decision" category is less common than "Upheld Decision," suggesting the majority of health plan decisions align with initial determinations.

Filters or Parameters Applied:

- Age ranges (0-10, 11-20, 21-30, etc.).
- Determination outcomes (e.g., overturned, upheld).



11_20
Determination Overturned Decision of Health Plan
Count 778

2. Count of Diagnosis and Treatment Summary by Report Year and Age Range

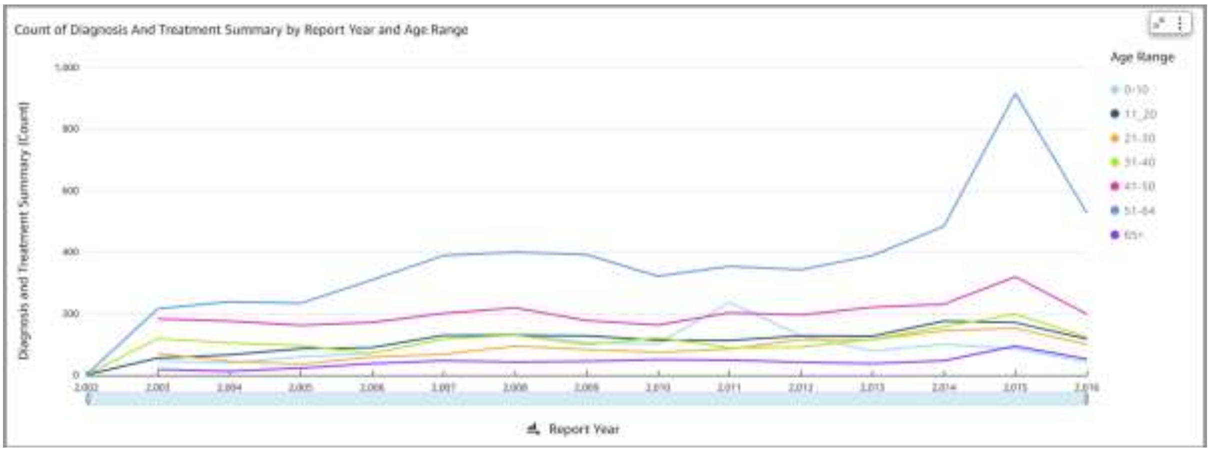
Purpose:
This line graph illustrates the number of diagnoses and treatments reported annually, categorized by age range.

Insights Derived:

- A steady increase in the number of diagnoses and treatments is observed over time, particularly in the 51-64 and 65+ age groups.
- Fluctuations in younger age ranges indicate inconsistent or lower levels of reporting.

Filters or Parameters Applied:

- Report years (2002-2016).
- Age groups (same as above).



3. Count of Determination by Patient Gender

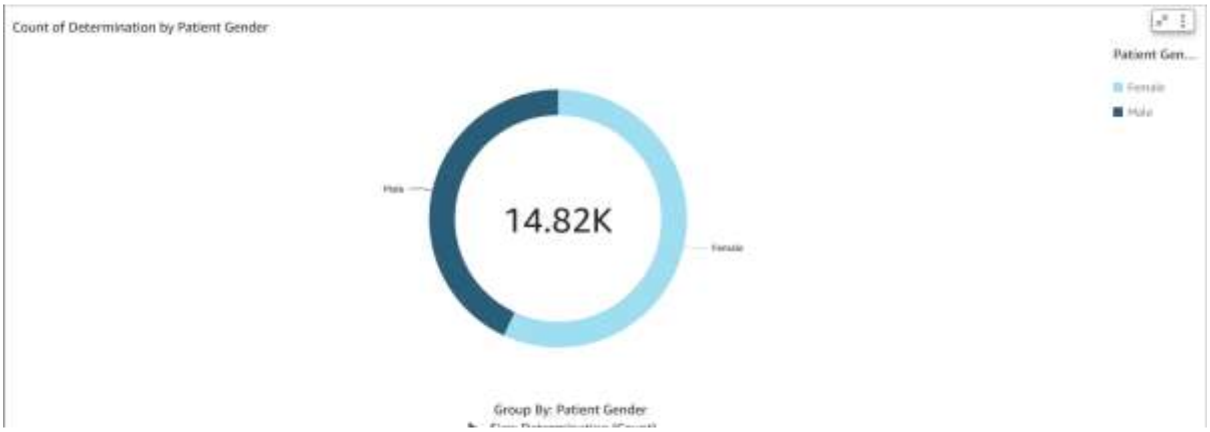
Purpose:
This pie chart highlights the proportion of health plan determinations for males versus females.

Insights Derived:

- A balanced distribution between male and female patients, slightly skewed toward females receiving more determinations.
- Potential gender-related trends or disparities in health plan review processes.

Filters or Parameters Applied:

- Patient gender (Male, Female).
- Determination counts.



4. Count of Records by Type

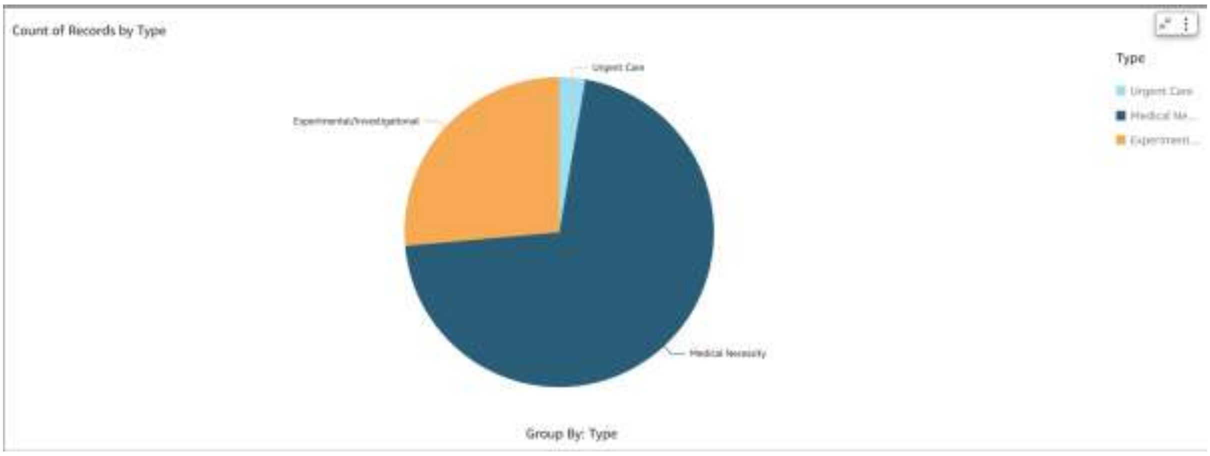
Purpose:
A horizontal bar chart showing the distribution of records based on the type of service (e.g., "Urgent Care," "Medical Necessity").

Insights Derived:

- "Medical Necessity" dominates the record count, reflecting its importance in health plan assessments.
- Urgent care records are significantly fewer, indicating lower frequencies of urgent cases under review.

Filters or Parameters Applied:

- Record types (e.g., "Urgent Care," "Medical Necessity")



5. Count of Determination by Patient Gender and Determination Outcome

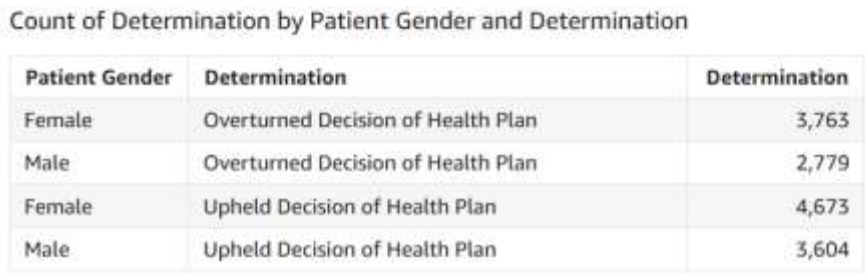
Purpose:
A clustered bar chart comparing determination outcomes ("Overturned" vs. "Upheld") across genders.

Insights Derived:

- Females show a higher count of "Upheld Decision" cases compared to males.
- Disparity in "Overturned Decisions" between genders might point to gender-based discrepancies in health plan appeals.

Filters or Parameters Applied:

- Determination type and patient gender.



Count of Determination by Patient Gender and Determination

Patient Gender	Determination	Determination
Female	Overturned Decision of Health Plan	3,763
Male	Overturned Decision of Health Plan	2,779
Female	Upheld Decision of Health Plan	4,673
Male	Upheld Decision of Health Plan	3,604

6. Count of Records by Determination and Type

Purpose

This bar chart categorizes health records by the type of service (e.g., "Medical Necessity," "Urgent Care," "Experimental Treatments") and the determination outcome ("Overturned" or "Upheld").

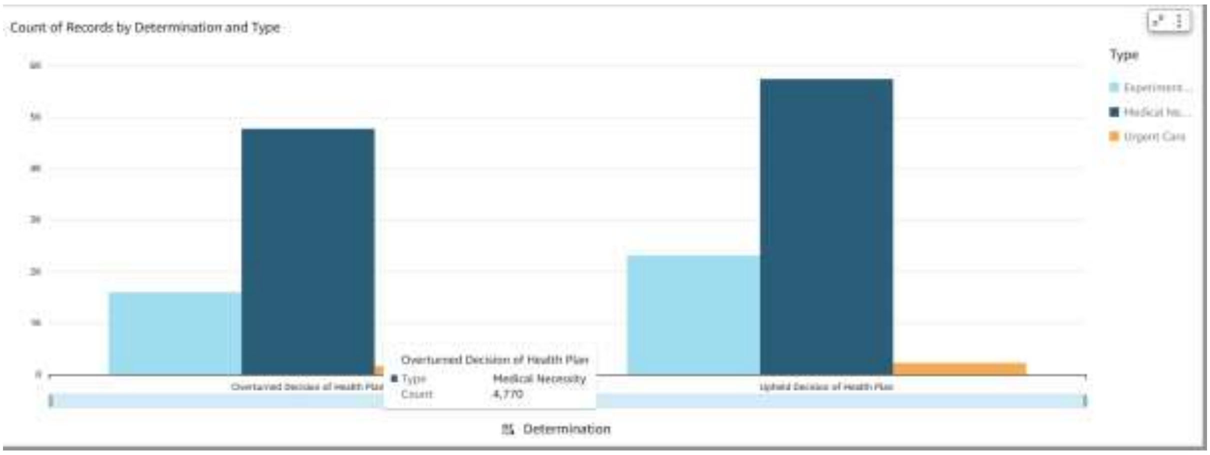
Insights Derived

- **Medical Necessity**: This category has the highest number of determinations, with most decisions being upheld. This suggests a higher degree of alignment with initial assessments for medical necessity cases.
- **Experimental Treatments**: A noticeable proportion of these cases are overturned, indicating potential challenges in meeting approval criteria for novel treatments.
- **Urgent Care**: Although a smaller category, urgent care records show a balanced distribution between overturned and upheld decisions.

Filters or Parameters Applied

- **Service Types**: Filtered by "Medical Necessity," "Urgent Care," and "Experimental Treatments."

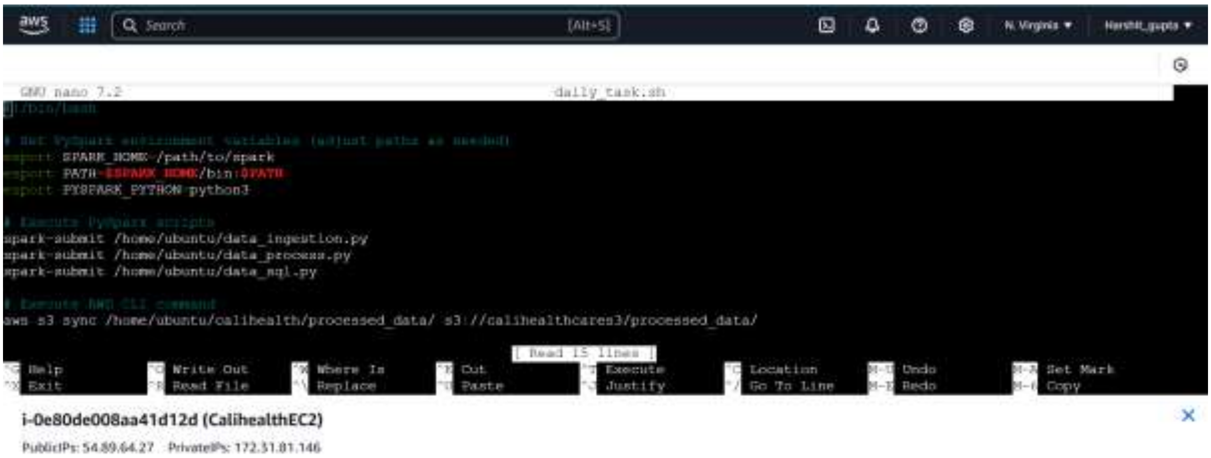
- **Determination Outcomes**: Segmented into "Overturned" and "Upheld."



5. Bonus Task: Automation of the Pipeline (5 Points)

Objective: Automate the entire data pipeline from ingestion to visualization:

1. **Develop an Automation Script (1 Point):** ○ Automate data retrieval, processing, and storage steps.

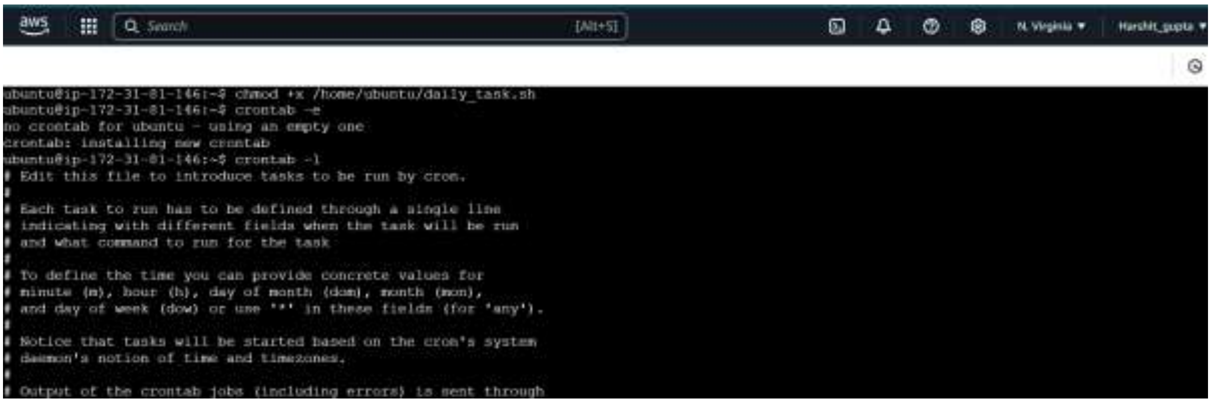


```
GNU nano 7.2                                          daily_task.sh
#!/bin/bash

# Set PySpark environment variables (adjust paths as needed)
export SPARK_HOME=/path/to/spark
export PATH=$SPARK_HOME/bin:$PATH
export PYSPARK_PYTHON=python3

# Execute PySpark scripts
spark-submit /home/ubuntu/data_ingestion.py
spark-submit /home/ubuntu/data_process.py
spark-submit /home/ubuntu/data_sql.py

# Execute AWS CLI command
aws s3 sync /home/ubuntu/calihealth/processed_data/ s3://calihealthcares3/processed_data/

                                              [ Read 15 lines ]
^G Help       ^O Write Out  ^W Where Is   ^K Cut        ^T Execute    ^C Location   M-U Undo      M-A Set Mark
^X Exit       ^R Read File  ^\ Replace    ^U Paste      ^J Justify    ^/ Go To Line M-E Redo      M-6 Copy
```

i-0e80de008aa41d12d (CalihealthEC2)

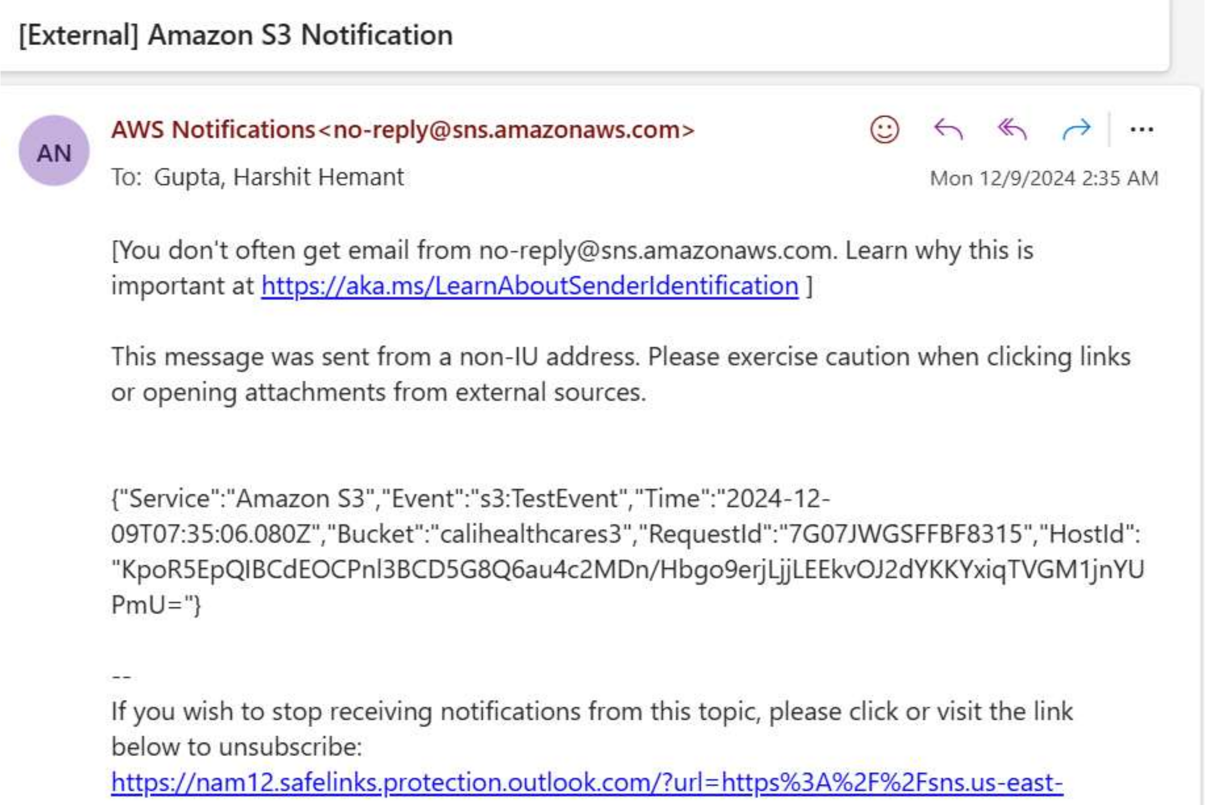
PublicIPs: 54.89.64.27 PrivateIPs: 172.31.81.146

2. **Set Up Scheduling Tools (1.5 Points):** ○ Use **cron jobs** or **AWS Lambda** to trigger the pipeline on a schedule or upon new data uploads to S3 bucket.



```
ubuntu@ip-172-31-81-146:~$ chmod +x /home/ubuntu/daily_task.sh
ubuntu@ip-172-31-81-146:~$ crontab -e
no crontab for ubuntu - using an empty one
crontab: installing new crontab
ubuntu@ip-172-31-81-146:~$ crontab -l
# Edit this file to introduce tasks to be run by cron.
#
# Each task to run has to be defined through a single line
# indicating with different fields when the task will be run
# and what command to run for the task
#
# To define the time you can provide concrete values for
# minute (m), hour (h), day of month (dom), month (mon),
# and day of week (dow) or use '*' in these fields (for 'any').
#
# Notice that tasks will be started based on the cron's system
# daemon's notion of time and timezones.
#
# Output of the crontab jobs (including errors) is sent through
```

3. **Implement Logging and Notifications (1.5 Points):** ○ Use **AWS SNS** or email to notify users of pipeline status (success or failure).



[External] Amazon S3 Notification

AN

AWS Notifications<no-reply@sns.amazonaws.com>

To: Gupta, Harshit Hemant

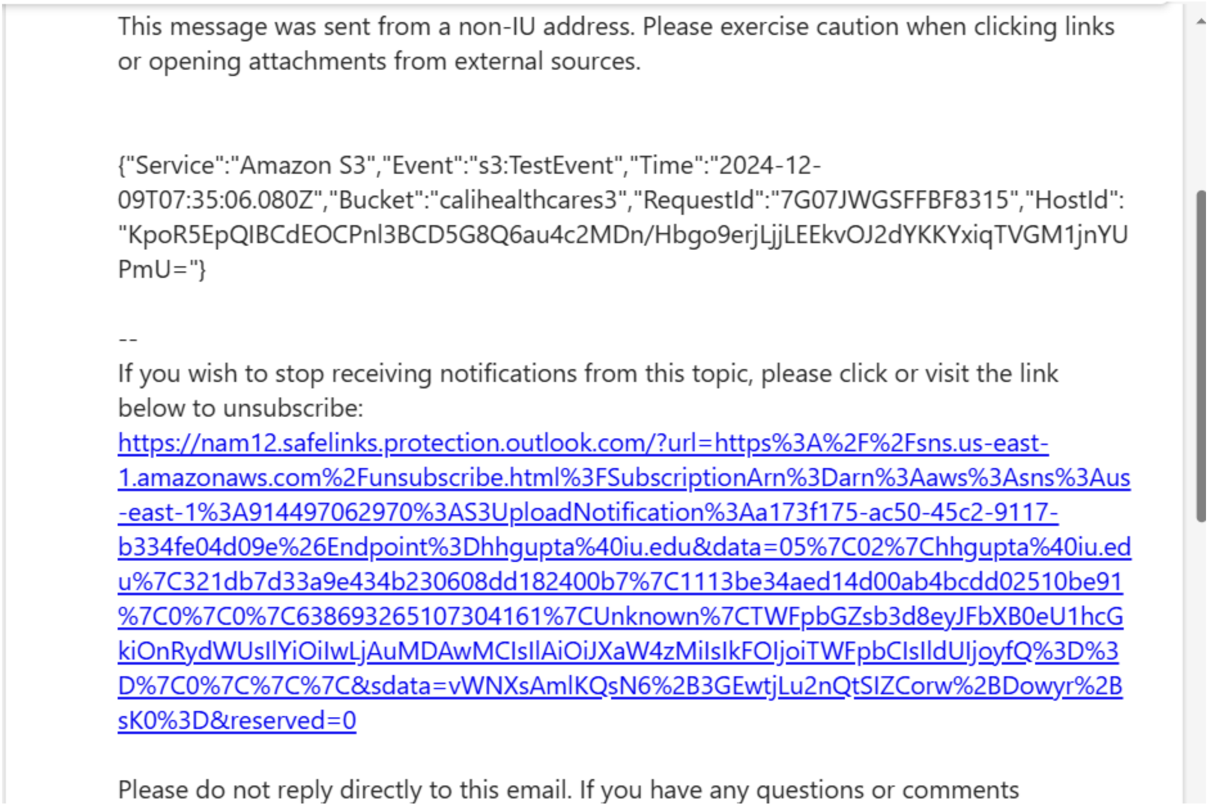
Mon 12/9/2024 2:35 AM

[You don't often get email from no-reply@sns.amazonaws.com. Learn why this is important at https://aka.ms/LearnAboutSenderIdentification]

This message was sent from a non-IU address. Please exercise caution when clicking links or opening attachments from external sources.

{"Service":"Amazon S3","Event":"s3:TestEvent","Time":"2024-12-09T07:35:06.080Z","Bucket":"calihealthcares3","RequestId":"7G07JWGSFFBF8315","HostId":"KpoR5EpQIBCdEOCPnl3BCD5G8Q6au4c2MDn/Hbgo9erjLjjLEEkvOJ2dYKKYxiqTVGM1jnYUPmU="}

--
If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsns.us-east-

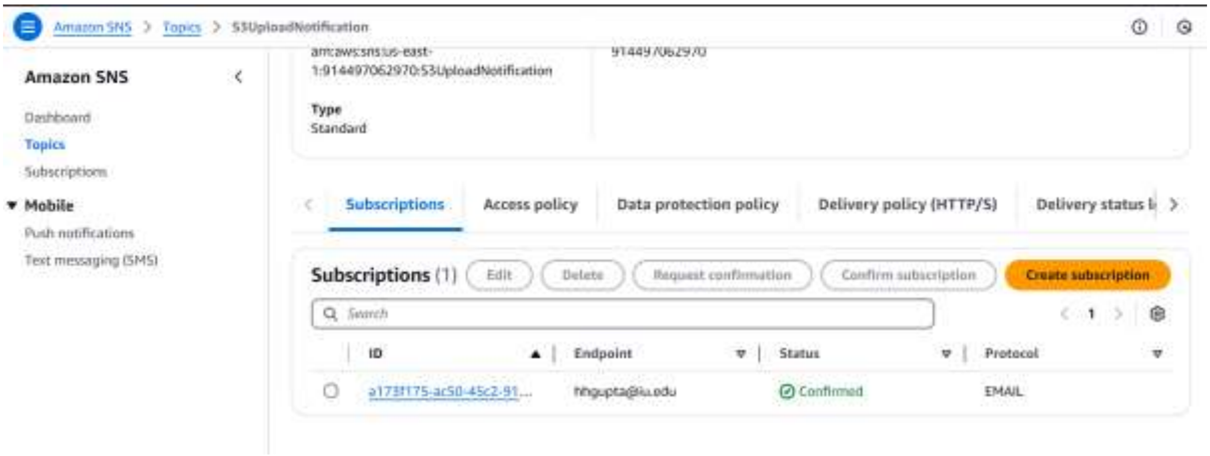


This message was sent from a non-IU address. Please exercise caution when clicking links or opening attachments from external sources.

{"Service":"Amazon S3","Event":"s3:TestEvent","Time":"2024-12-09T07:35:06.080Z","Bucket":"calihealthcares3","RequestId":"7G07JWGSFFBF8315","HostId":"KpoR5EpQIBCdEOCPnl3BCD5G8Q6au4c2MDn/Hbgo9erjLjjLEEkvOJ2dYKKYxiqTVGM1jnYUPmU="}

--
If you wish to stop receiving notifications from this topic, please click or visit the link below to unsubscribe:
https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fsns.us-east-1.amazonaws.com%2Funsubscribe.html%3FSubscriptionArn%3Darn%3Aaws%3Asns%3Aus-east-1%3A914497062970%3AS3UploadNotification%3Aa173f175-ac50-45c2-9117-b334fe04d09e%26Endpoint%3Dhhgupta%40iu.edu&data=05%7C02%7Chhgupta%40iu.edu%7C321db7d33a9e434b230608dd182400b7%7C1113be34aed14d00ab4bcdd02510be91%7C0%7C0%7C638693265107304161%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=vWNXsAmlKQsN6%2B3GEwtjLu2nQtSIZCorw%2BDowyr%2BsK0%3D&reserved=0

Please do not reply directly to this email. If you have any questions or comments



arn:aws:sns:us-east-1:914497062970:S3UploadNotification

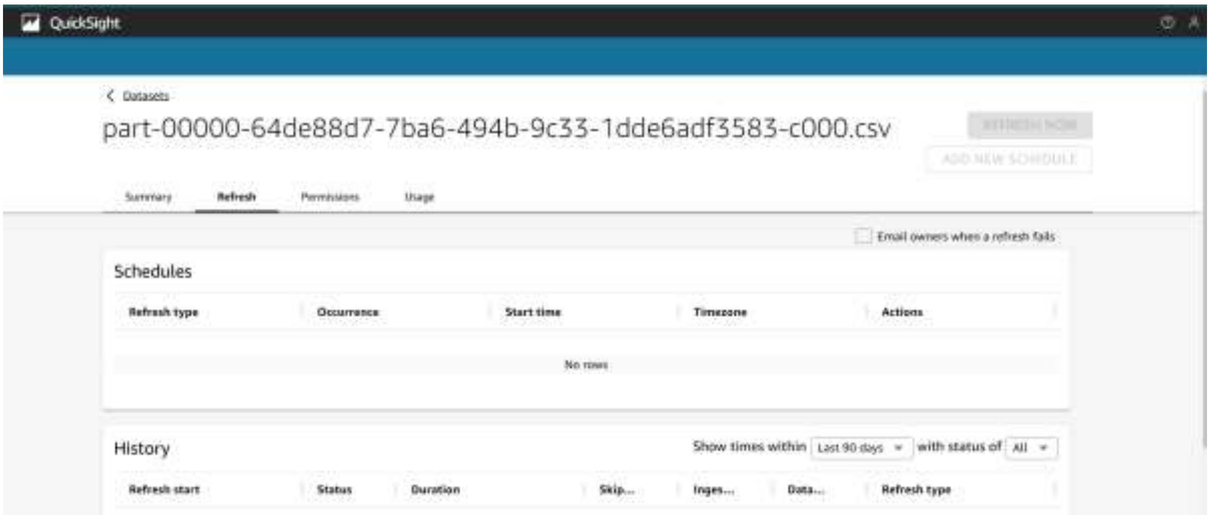
914497062970

Type
Standard

Subscriptions (1)

	ID	Endpoint	Status	Protocol
○	a173f175-ac50-45c2-91...	hhgupta@iu.edu	Confirmed	EMAIL

4. **Automate Dashboard Updates (1 Point):** ○ Configure QuickSight or Power BI to refresh data periodically for real-time insights.



QuickSight

< Datasets

part-00000-64de88d7-7ba6-494b-9c33-1dde6adf3583-c000.csv

REFRESH NOW

ADD NEW SCHEDULE

Summary | Refresh | Permissions | Usage

□ Email owners when a refresh fails

Schedules

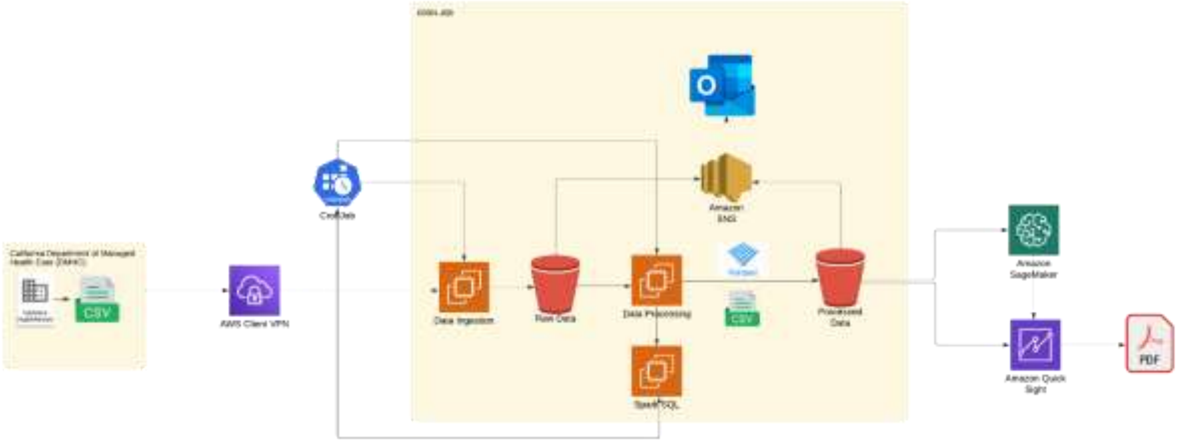
Refresh type	Occurrence	Start time	Timezone	Actions
No rows				

History

Show times within Last 90 days ▾ with status of All ▾

Refresh start	Status	Duration	Skip...	Inges...	Data...	Refresh type

Bonus: Submit an architecture diagram showing the entire pipeline.


California Department of Managed Health Care (DMHC)
CSV
AWS Client VPN
CronJob
Data Ingestion
Raw Data
Data Processing
Spark SQL
Amazon SNS
CSV
Processed Data
Amazon SageMaker
Amazon Quick Sight
PDF