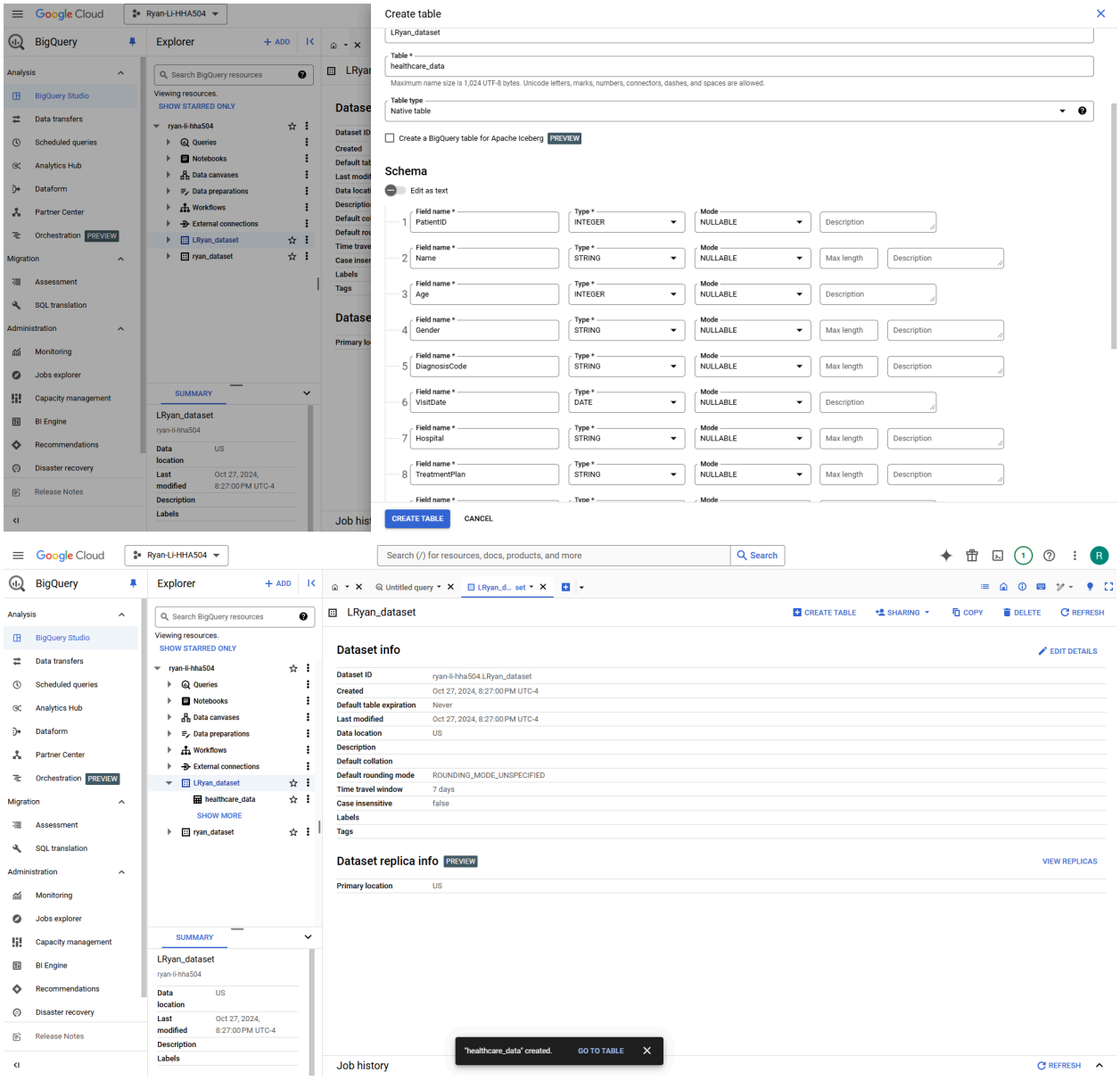
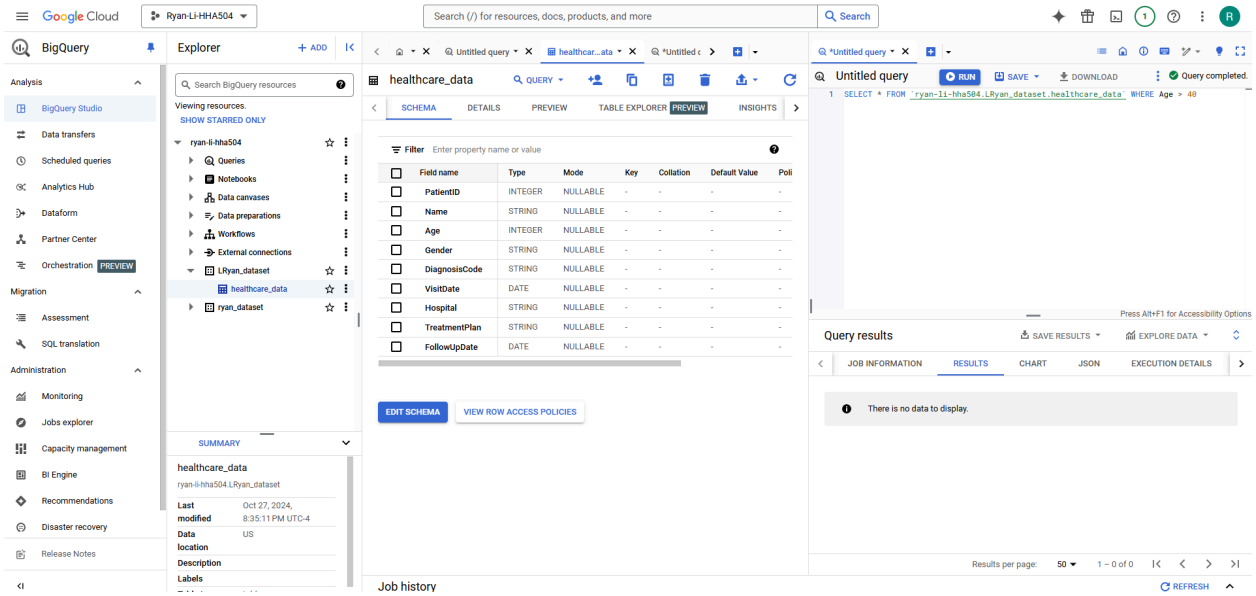
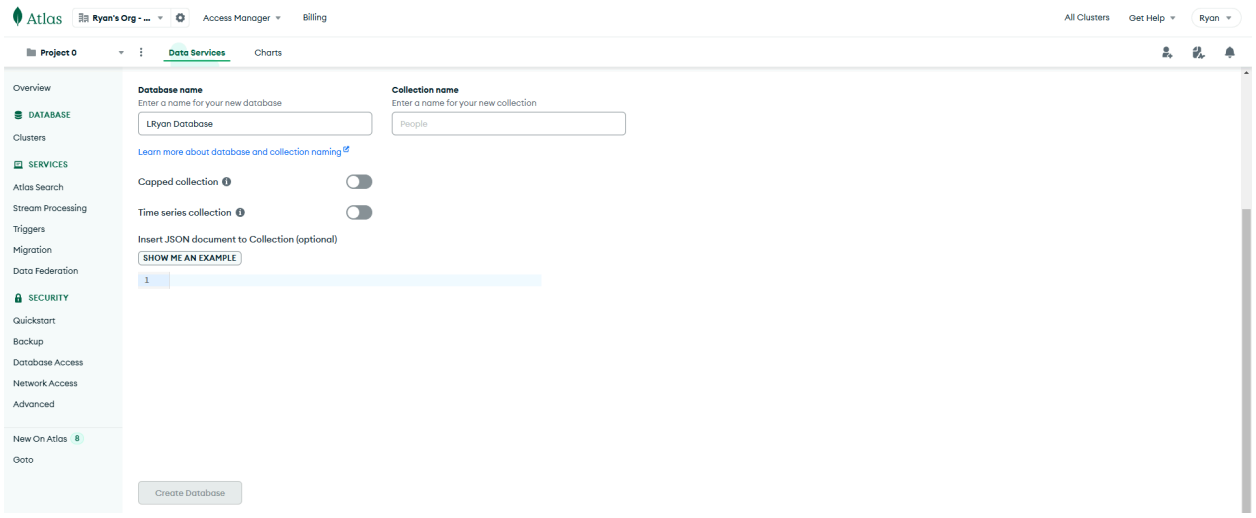

Google Cloud
Ryan-LI-HHA504
BigQuery
Explorer
+ ADD
Search BigQuery resources
Viewing resources.
SHOW STARRED ONLY
ryan-li-hha504
Queries
Notebooks
Data canvases
Data preparations
Workflows
External connections
LRyan_dataset
ryan_dataset
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Orchestration PREVIEW
Migration
Assessment
SQL translation
Administration
Monitoring
Jobs explorer
Capacity management
BI Engine
Recommendations
Disaster recovery
Release Notes
SUMMARY
LRyan_dataset
ryan-li-hha504
Data location US
Last modified Oct 27, 2024, 8:27:00 PM UTC-4
Description
Labels
Create table
LRyan_dataset
Table * healthcare_data
Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.
Table type Native table
Create a BigQuery table for Apache Iceberg PREVIEW
Schema
Edit as text
1 Field name * PatientID Type * INTEGER Mode NULLABLE Description
2 Field name * Name Type * STRING Mode NULLABLE Max length Description
3 Field name * Age Type * INTEGER Mode NULLABLE Description
4 Field name * Gender Type * STRING Mode NULLABLE Max length Description
5 Field name * DiagnosisCode Type * STRING Mode NULLABLE Max length Description
6 Field name * VisitDate Type * DATE Mode NULLABLE Description
7 Field name * Hospital Type * STRING Mode NULLABLE Max length Description
8 Field name * TreatmentPlan Type * STRING Mode NULLABLE Max length Description
CREATE TABLE
CANCEL

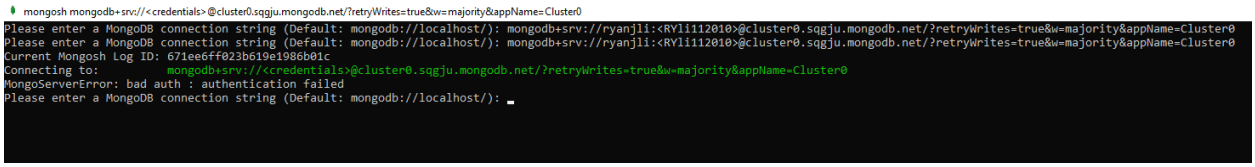
Google Cloud
Ryan-LI-HHA504
Search (/) for resources, docs, products, and more
Search
BigQuery
Explorer
+ ADD
Search BigQuery resources
Viewing resources.
SHOW STARRED ONLY
ryan-li-hha504
Queries
Notebooks
Data canvases
Data preparations
Workflows
External connections
LRyan_dataset
healthcare_data
SHOW MORE
ryan_dataset
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Orchestration PREVIEW
Migration
Assessment
SQL translation
Administration
Monitoring
Jobs explorer
Capacity management
BI Engine
Recommendations
Disaster recovery
Release Notes
SUMMARY
LRyan_dataset
ryan-li-hha504
Data location US
Last modified Oct 27, 2024, 8:27:00 PM UTC-4
Description
Labels
Untitled query
LRyan_d... set
LRyan_dataset
CREATE TABLE
SHARING
COPY
DELETE
REFRESH
Dataset info
EDIT DETAILS
Dataset ID ryan-li-hha504.LRyan_dataset
Created Oct 27, 2024, 8:27:00 PM UTC-4
Default table expiration Never
Last modified Oct 27, 2024, 8:27:00 PM UTC-4
Data location US
Description
Default collation
Default rounding mode ROUNDING_MODE_UNSPECIFIED
Time travel window 7 days
Case insensitive false
Labels
Tags
Dataset replica info PREVIEW
VIEW REPLICAS
Primary location US
"healthcare_data" created.
GO TO TABLE
Job history
REFRESH

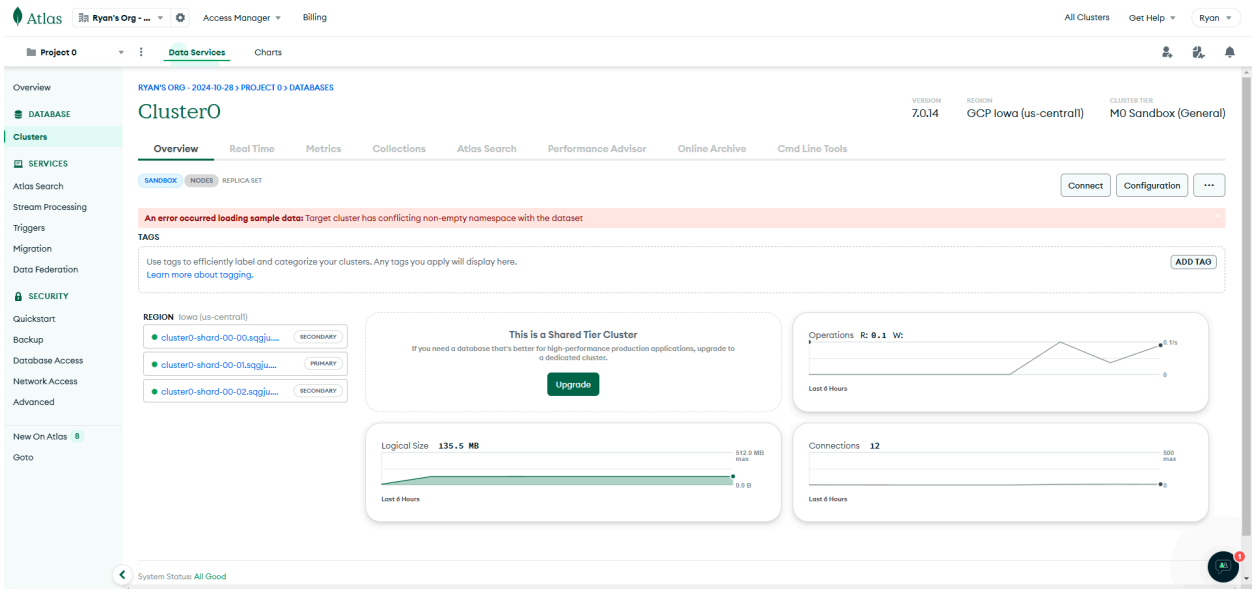

Google Cloud
Ryan-Li-HHA504
Search (/) for resources, docs, products, and more
Search
BigQuery
Explorer
+ ADD
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Orchestration
PREVIEW
Migration
Assessment
SQL translation
Administration
Monitoring
Jobs explorer
Capacity management
BI Engine
Recommendations
Disaster recovery
Release Notes
Search BigQuery resources
Viewing resources.
SHOW STARRED ONLY
ryan-li-hha504
Queries
Notebooks
Data canvases
Data preparations
Workflows
External connections
LRyan_dataset
healthcare_data
ryan_dataset
SUMMARY
healthcare_data
ryan-li-hha504.LRyan_dataset
Last modified
Oct 27, 2024, 8:35:11 PM UTC-4
Data location
US
Description
Labels
healthcare_data
QUERY
SCHEMA
DETAILS
PREVIEW
TABLE EXPLORER
PREVIEW
INSIGHTS
Filter
Enter property name or value
Field name
Type
Mode
Key
Collation
Default Value
Poli
PatientID
INTEGER
NULLABLE
Name
STRING
NULLABLE
Age
INTEGER
NULLABLE
Gender
STRING
NULLABLE
DiagnosisCode
STRING
NULLABLE
VisitDate
DATE
NULLABLE
Hospital
STRING
NULLABLE
TreatmentPlan
STRING
NULLABLE
FollowUpDate
DATE
NULLABLE
EDIT SCHEMA
VIEW ROW ACCESS POLICIES
Job history
Untitled query
RUN
SAVE
DOWNLOAD
Query completed.
1 SELECT * FROM `ryan-li-hha504.LRyan_dataset.healthcare_data` WHERE Age > 40
Press Alt+F1 for Accessibility Options
Query results
SAVE RESULTS
EXPLORE DATA
JOB INFORMATION
RESULTS
CHART
JSON
EXECUTION DETAILS
There is no data to display.
Results per page:
50
1 – 0 of 0
REFRESH

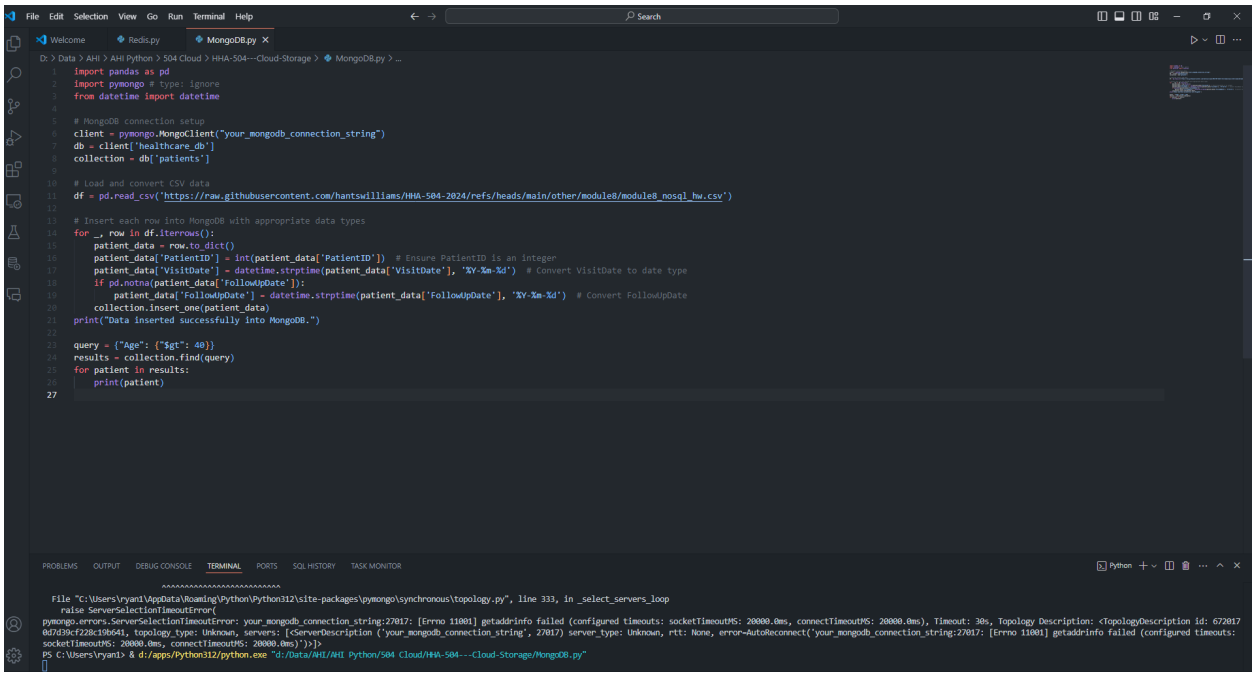
MongoDB


Atlas
Ryan's Org - ...
Access Manager
Billing
All Clusters
Get Help
Ryan
Project 0
Data Services
Charts
Overview
DATABASE
Clusters
SERVICES
Atlas Search
Stream Processing
Triggers
Migration
Data Federation
SECURITY
Quickstart
Backup
Database Access
Network Access
Advanced
New On Atlas
Goto
Database name
Enter a name for your new database
LRyan Database
Collection name
Enter a name for your new collection
People
Learn more about database and collection naming
Capped collection
Time series collection
Insert JSON document to Collection (optional)
SHOW ME AN EXAMPLE
1
Create Database


Select C:\Users\ryan1\AppData\Local\Temp\dtemp-cbdffa6836563406-60.dop\mongosh-2.3.2-win32-x64\bin\mongosh.exe
Please enter a MongoDB connection string (Default: mongodb://localhost/): mongosh "mongodb+srv://cluster0.sqgju.mongodb.net/" --apiVersion 1 --username ryanjli
Please enter a MongoDB connection string (Default: mongodb://localhost/): mongosh "mongodb+srv://cluster0.sqgju.mongodb.net/" --apiVersion 1 --username ryanjlimongosh "mongodb+srv://cluster0.sqgju.mongodb.net/" --apiVersion 1 --username ryanjli
MongoshInvalidInputError: [COMMON-10001] Invalid URI: mongosh "mongodb+srv://cluster0.sqgju.mongodb.net/" --apiVersion 1 --username ryanjlimongosh "mongodb+srv://cluster0.sqgju.mongodb.net/" --apiVersion 1 --username ryanjli
Please enter a MongoDB connection string (Default: mongodb://localhost/):

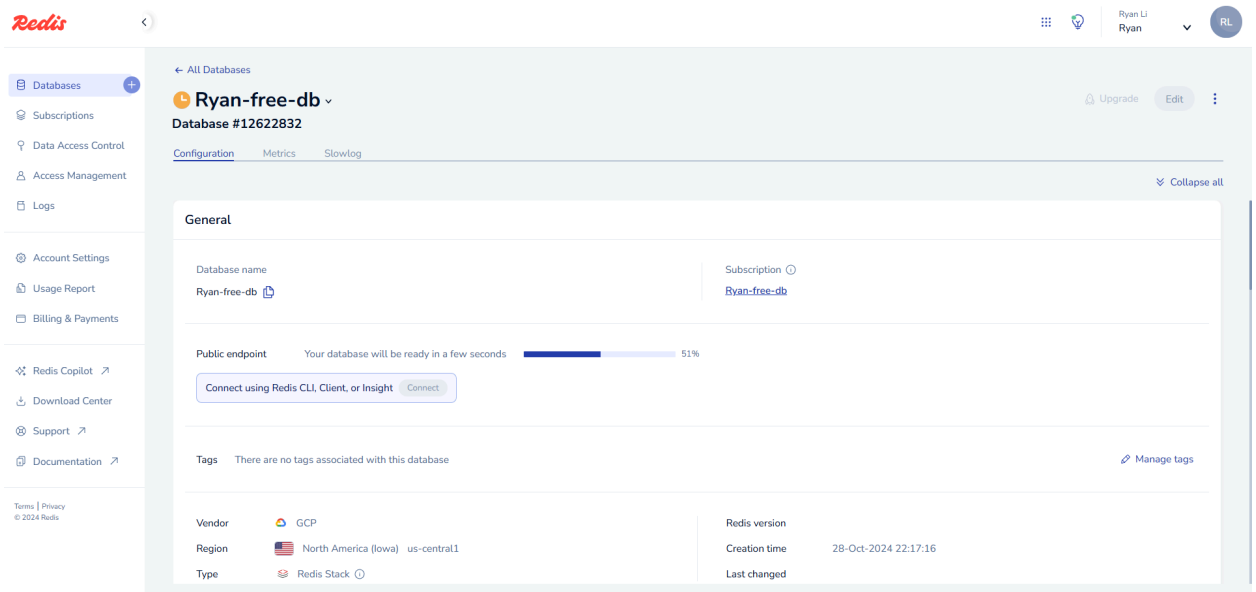

mongosh mongodb+srv://<credentials>@cluster0.sqgju.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0
Please enter a MongoDB connection string (Default: mongodb://localhost/): mongodb+srv://ryanjli:<RYli112010>@cluster0.sqgju.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0
Please enter a MongoDB connection string (Default: mongodb://localhost/): mongodb+srv://ryanjli:<RYli112010>@cluster0.sqgju.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0
Current Mongosh Log ID: 671ee6ff023b619e1986b01c
Connecting to: mongodb+srv://<credentials>@cluster0.sqgju.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0
MongoServerError: bad auth : authentication failed
Please enter a MongoDB connection string (Default: mongodb://localhost/):

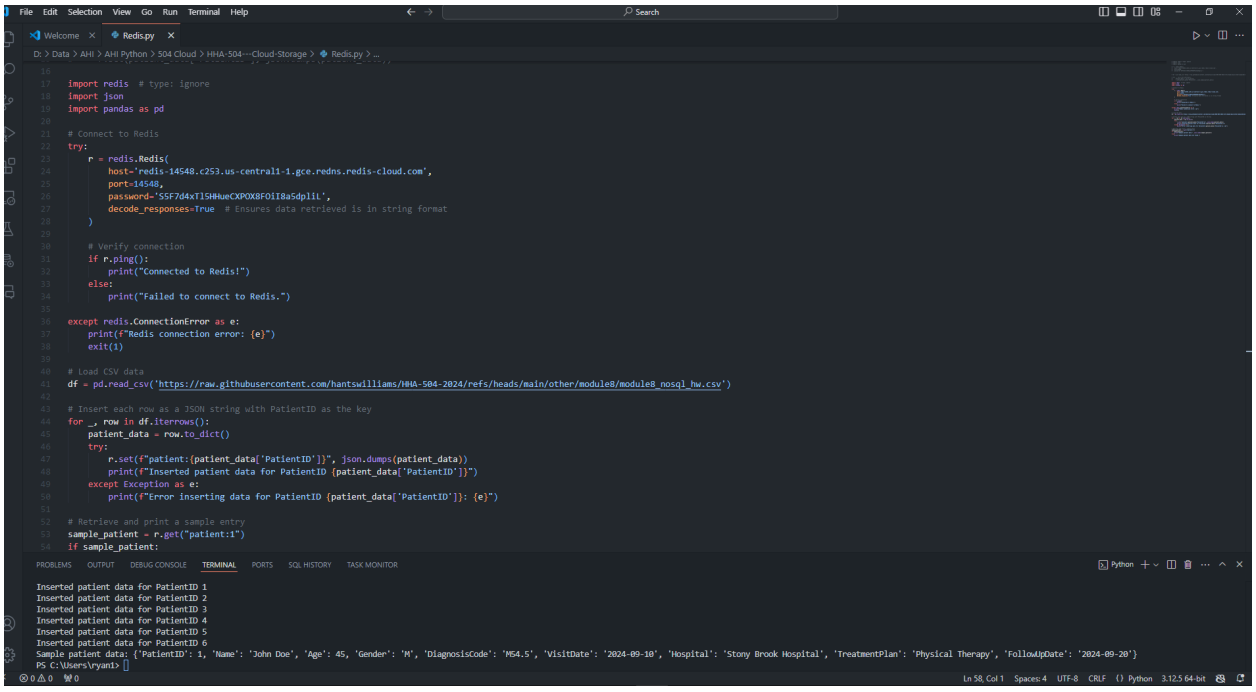

Atlas
Ryan's Org - ...
Access Manager
Billing
All Clusters
Get Help
Ryan
Project 0
Data Services
Charts
Overview
DATABASE
Clusters
SERVICES
Atlas Search
Stream Processing
Triggers
Migration
Data Federation
SECURITY
Quickstart
Backup
Database Access
Network Access
Advanced
New On Atlas
Goto
RYAN'S ORG - 2024-10-28 > PROJECT 0 > DATABASES
Cluster0
VERSION 7.0.14
REGION GCP Iowa (us-central1)
CLUSTER TIER M0 Sandbox (General)
Overview
Real Time
Metrics
Collections
Atlas Search
Performance Advisor
Online Archive
Cmd Line Tools
SANDBOX
NODES
REPLICA SET
Connect
Configuration
An error occurred loading sample data: Target cluster has conflicting non-empty namespace with the dataset
TAGS
Use tags to efficiently label and categorize your clusters. Any tags you apply will display here.
Learn more about tagging.
ADD TAG
REGION Iowa (us-central1)
cluster0-shard-00-00.sqgju.....
SECONDARY
cluster0-shard-00-01.sqgju.....
PRIMARY
cluster0-shard-00-02.sqgju.....
SECONDARY
This is a Shared Tier Cluster
If you need a database that's better for high-performance production applications, upgrade to a dedicated cluster.
Upgrade
Operations R: 0.1 W:
Last 6 Hours
Logical Size 135.5 MB
Last 6 Hours
Connections 12
Last 6 Hours
System Status: All Good


File Edit Selection View Go Run Terminal Help
Search
Welcome
Redis.py
MongoDB.py
D: > Data > AHI > AHI Python > 504 Cloud > HHA-504---Cloud-Storage > MongoDB.py > ...
import pandas as pd
import pymongo # type: ignore
from datetime import datetime
# MongoDB connection setup
client = pymongo.MongoClient("your_mongodb_connection_string")
db = client['healthcare_db']
collection = db['patients']
# Load and convert CSV data
df = pd.read_csv('https://raw.githubusercontent.com/hantswilliams/HHA-504-2024/refs/heads/main/other/module8/module8_nosql_hw.csv')
# Insert each row into MongoDB with appropriate data types
for _, row in df.iterrows():
patient_data = row.to_dict()
patient_data['PatientID'] = int(patient_data['PatientID']) # Ensure PatientID is an integer
patient_data['VisitDate'] = datetime.strptime(patient_data['VisitDate'], '%Y-%m-%d') # Convert VisitDate to date type
if pd.notna(patient_data['FollowUpDate']):
patient_data['FollowUpDate'] = datetime.strptime(patient_data['FollowUpDate'], '%Y-%m-%d') # Convert FollowUpDate
collection.insert_one(patient_data)
print("Data inserted successfully into MongoDB.")
query = {"Age": {"$gt": 40}}
results = collection.find(query)
for patient in results:
print(patient)
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
SQL HISTORY
TASK MONITOR
Python
^^^^^^^^^^^^^^^^^^^^^^^^^^
File "C:\Users\ryan1\AppData\Roaming\Python\Python312\site-packages\pymongo\synchronous\topology.py", line 333, in _select_servers_loop
raise ServerSelectionTimeoutError(
pymongo.errors.ServerSelectionTimeoutError: your_mongodb_connection_string:27017: [Errno 11001] getaddrinfo failed (configured timeouts: socketTimeoutMS: 20000.0ms, connectTimeoutMS: 20000.0ms), Timeout: 30s, Topology Description: <TopologyDescription id: 672017
0d7d39cf228c19b641, topology_type: Unknown, servers: [<ServerDescription ('your_mongodb_connection_string', 27017) server_type: Unknown, rtt: None, error=AutoReconnect('your_mongodb_connection_string:27017: [Errno 11001] getaddrinfo failed (configured timeouts:
socketTimeoutMS: 20000.0ms, connectTimeoutMS: 20000.0ms)')>]>
PS C:\Users\ryan1> & d:/apps/Python312/python.exe "d:/Data/AHI/AHI Python/504 Cloud/HHA-504---Cloud-Storage/MongoDB.py"

Had difficulty importing the dataset into mongoDB. I tried using the shell to import the csv file locally but kept facing an error that made application close down. I also tried manually importing a json file into the database but kept receiving an error.

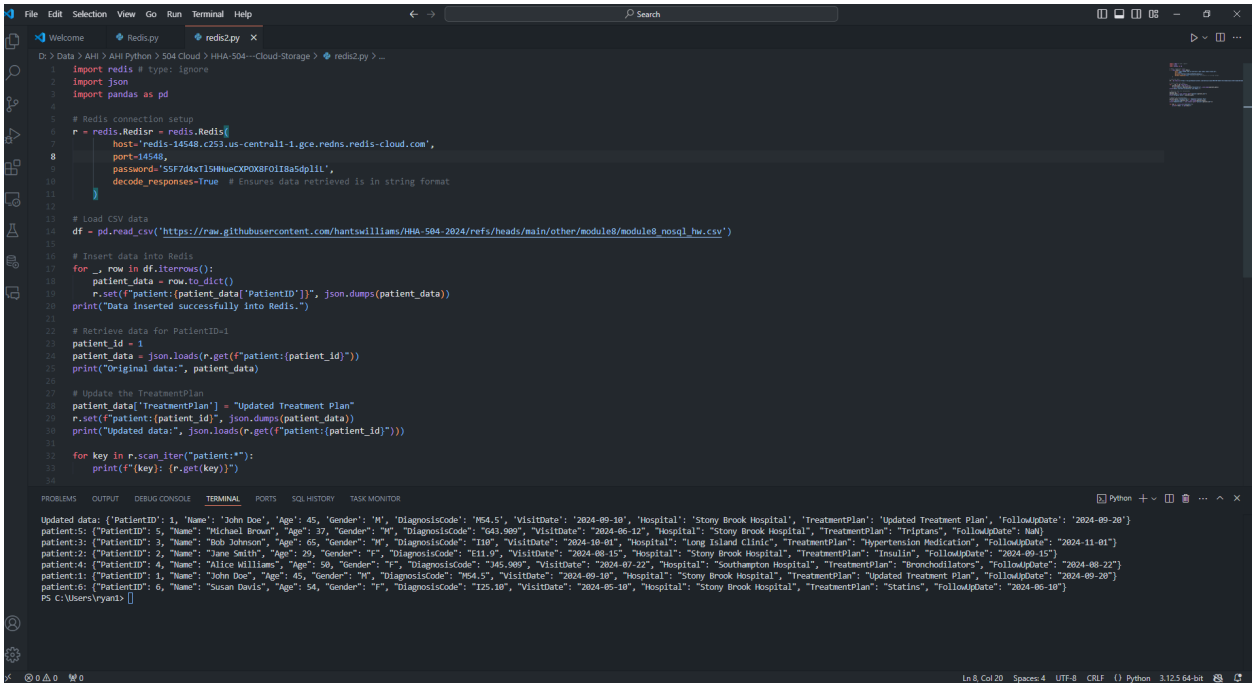
Redis





```python
# Insert each row as a JSON string with PatientID as the key
for _, row in df.iterrows():
    patient_data = row.to_dict()
    try:
        r.set(f"patient:{patient_data['PatientID']}", json.dumps(patient_data))
        print(f"Inserted patient data for PatientID {patient_data['PatientID']}")
    except Exception as e:
        print(f"Error inserting data for PatientID {patient_data['PatientID']}: {e}")

# Retrieve and print a sample entry
sample_patient = r.get("patient:1")
if sample_patient:
    print("Sample patient data:", json.loads(sample_patient))
else:
    print("Sample patient data not found.")
```



BigQuery

- Pretty simple process
- Create new dataset within bigquery on GCP
- Uploaded the provided data as a table manually within the bigquery dataset
- Ran the SQL query on a dataset to retrieve patients over the age of 40.
- Monitoring tools are useful to view resource usage and cost estimates
- Clear structured layout for datasets, tables and queries

MongoDB

- Created a mongoDB account and set up a new cluster
- Created new database and collection to store the healthcare data
- I used a python script using pymongo to read the csv file and convert each row to JSOn format
- I used the query functionality to retrieve records where the patients age was over 40
- Pretty user friendly and eas to set up
- FIxible document based approach

Redis

- Created new redis cloud account and new database instance
- I made a python script using the redis library
- Retrieved and updated data by querying patient records based on their ids
- Very minimal and clean interface
- Key value storage
- Efficient and simple fast data retrieval but best suited for scenarios with key value pairs data