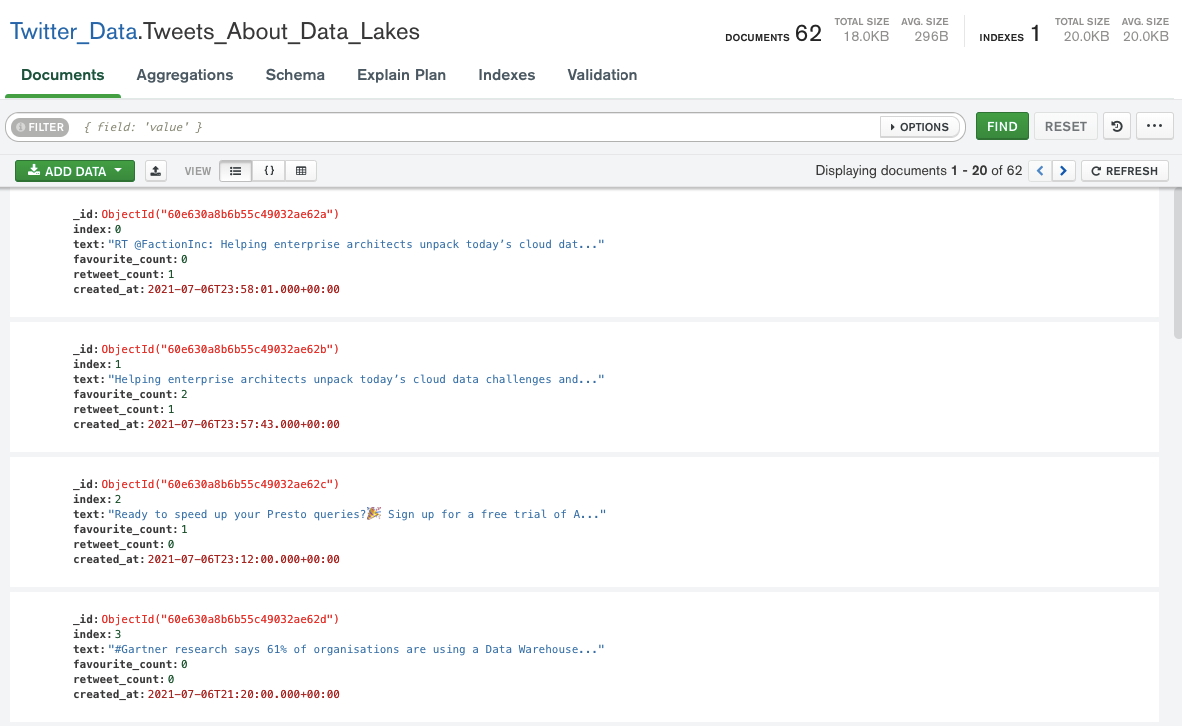
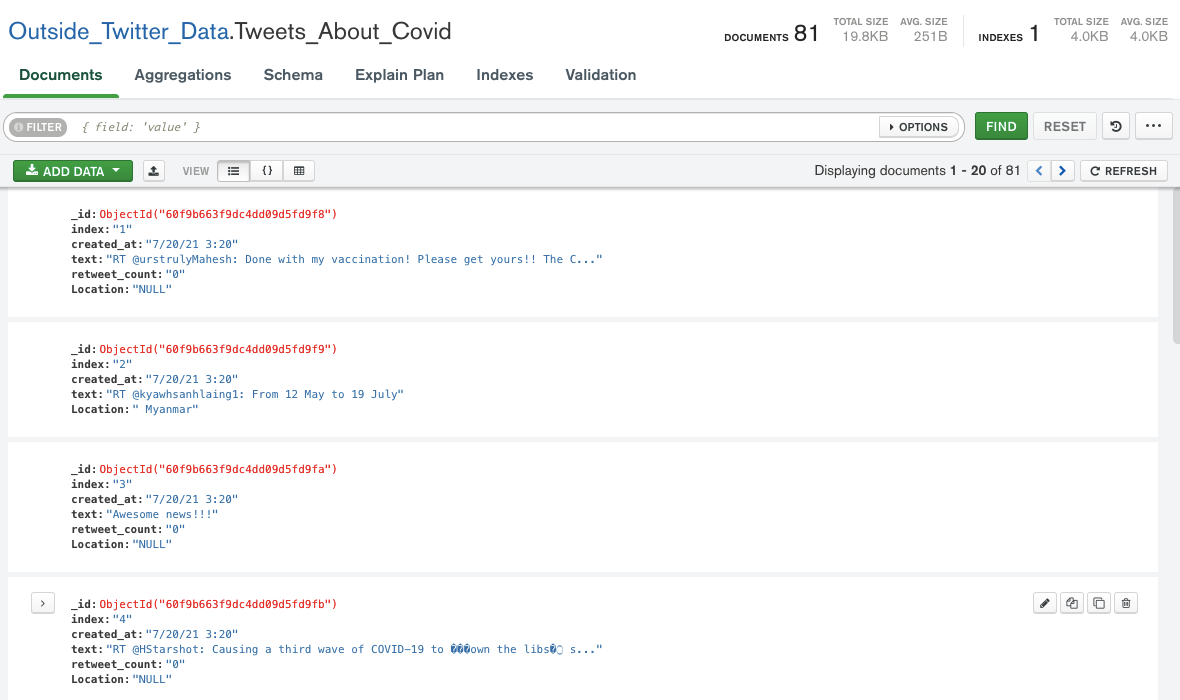
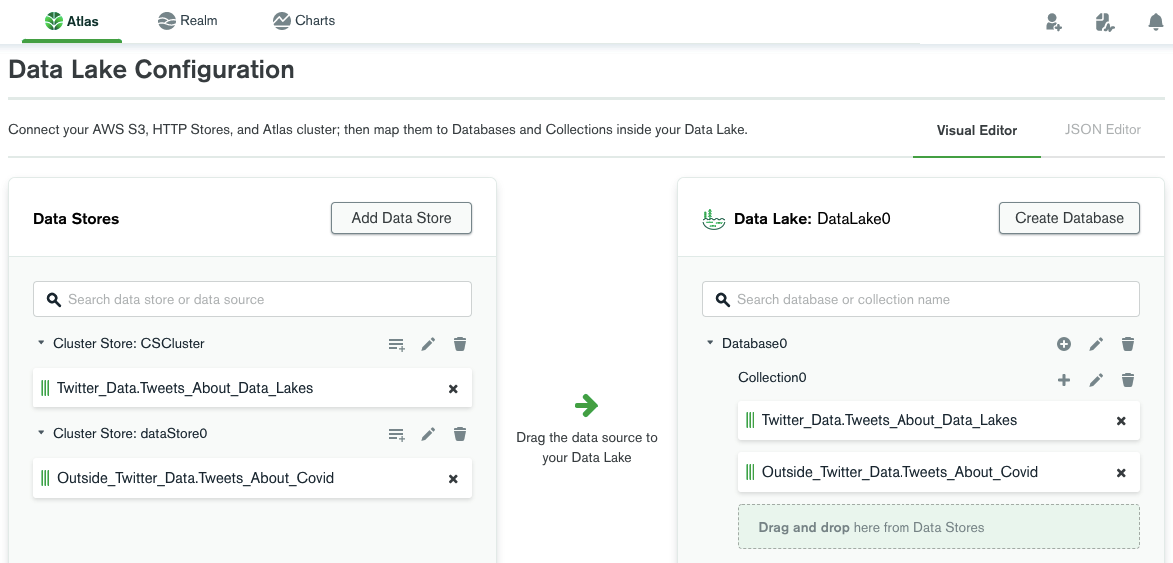
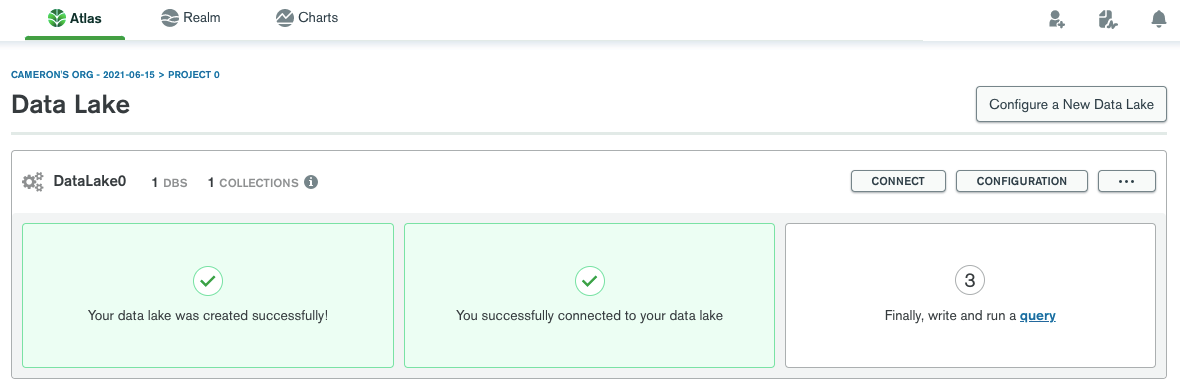
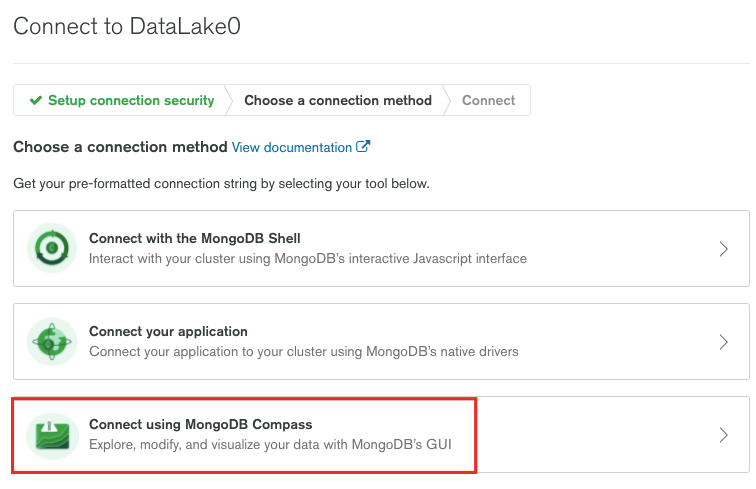
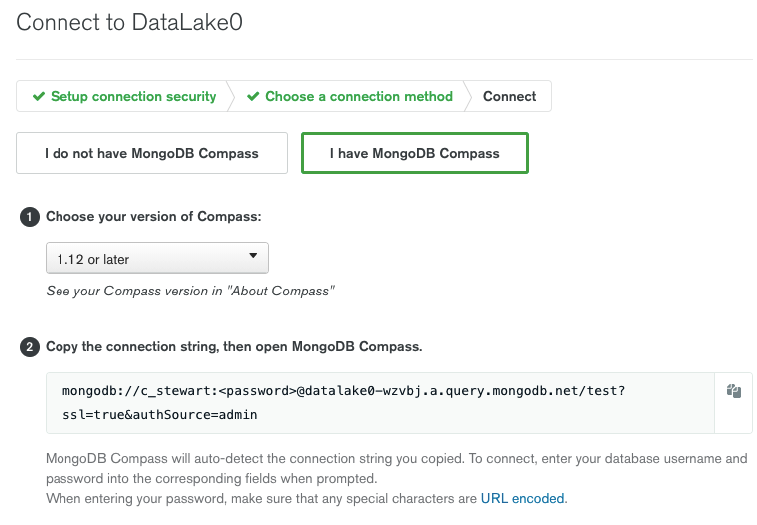
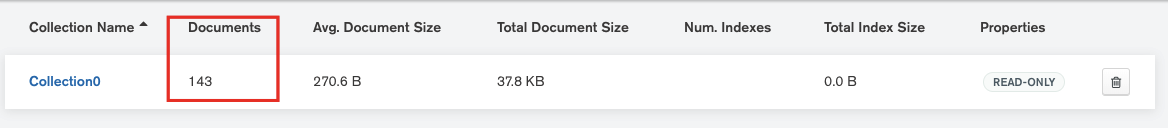
Data Lake Walkthrough

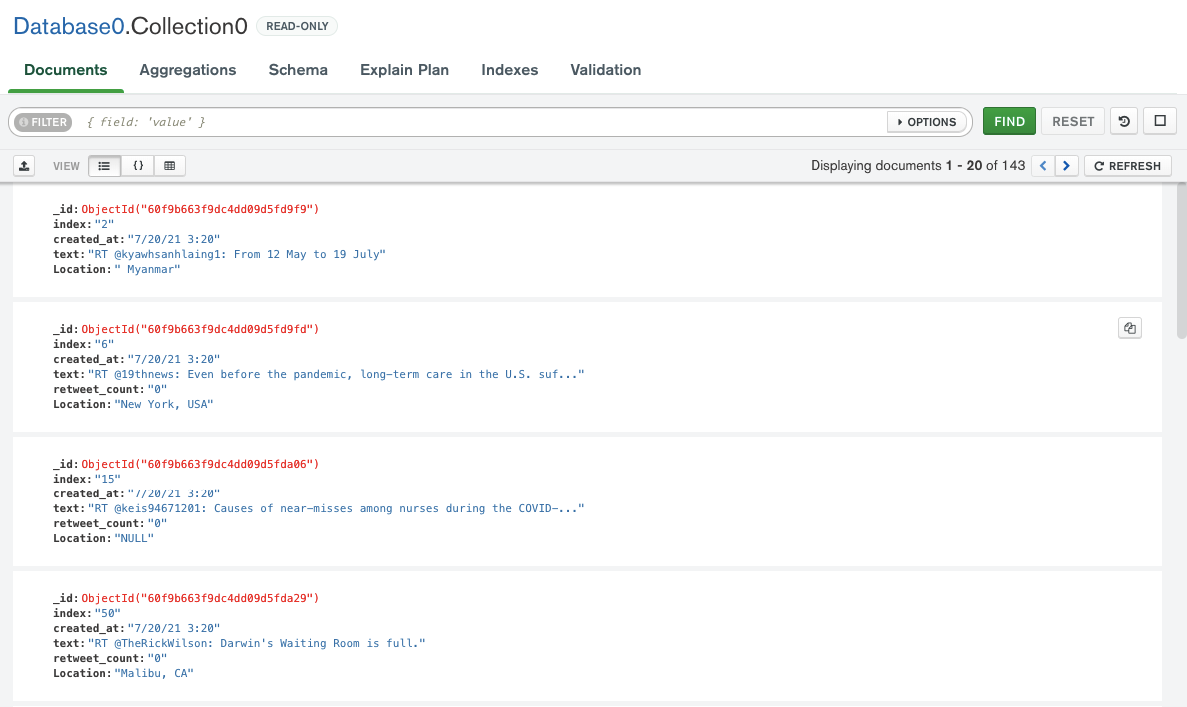
* 62 Tweets with the keyword ‘DataLakes’ were pulled directly into MongoDB using Python
  + Reference Jupyter Notebook for code used:
    - <https://github.com/nedeinlein/DBMFinal_Paper/blob/main/Gather_Tweets_and_Upload_to_MongoDB.ipynb>
  + References used to learn how to pull Twitter Data and push to MongoDB:
    - <https://github.com/alod83/data-science/blob/master/DataCollection/Twitter/get_tweets.ipynb>
    - https://medium.com/analytics-vidhya/how-to-upload-a-pandas-dataframe-to-mongodb-ffa18c0953c1
* Pulled in tweets about Data Lakes were placed in a database called 'Twitter\_Data'. The collection is called ‘Tweets\_About\_Data\_Lakes’. Notice there are 62 documents.
* 
* Pulled in tweets about Covid from our Relational Database data by importing a csv into database called 'Outside\_Twitter\_Data'. The collection is called ‘Tweets\_About\_Covid’.
* 

* We used the Data Lake feature in MongoDB to set up the Data Lake
  + To configure the Data Lake, just drag any collection from any database into the Data Lake
    - The Data Lake can hold collections from multiple databases
      * This allows for data to be easily queried across databases. This would be much more complex NoSQL or Relational Databases without a Data Lake.
    - The Data Lake can continuously take in new data and allows for data to be unstructured
    - In the example of configuring a Data Lake below, we pulled in the two collections created above which are stored in two different databases
  + 

* Once your Data Lake is properly configured, you will receive confirmation shown below and we can begin querying data

* 
* You can connect to your Data Lake using the same methods in MongoDB as a standard database. We chose to use MongoDB Compass.

* 
* 
* In MongoDB Compass, we can see the Data Lake took the collections from the two databases and combined them into one collection with 143 documents. ‘Database0’ is the Data Lake and ‘Collection0’ is the combined collection.
  + 

* 
* Access Datalake using MongoShell
  + use Database0
  + 'switched to db Database0'

* Query for documents with the string " data " AND a retweet count of 0 resulted in one document being returned from each of the original data sets which are combined in the Data Lake
  + db.Collection0.find({"retweet\_count":0,"text":/ data /}).sort({created\_at: -1}).pretty()
  + { \_id: ObjectId("60f9e95fffc0381c73f2c993"),  
     index: '35',  
     created\_at: 2021-07-20T10:20:00.000Z,  
     text: 'RT @SuicidePrevAU: The @aihw released updated data on suicide &amp; self-harm monitoring today. Despite increased calls to support lines &amp; ment \_',  
     retweet\_count: 0,  
     Location: 'Canberra, Australia' }

{ \_id: ObjectId("60e630a8b6b55c49032ae632"),  
 index: 8,  
 text: 'To efficiently ingest data into a data lake, your team must properly evaluate the environment and technology choices. NEOS shares its lessons learned. \nhttps://t.co/V474QDX1LP\n#datalake #database #insurance <https://t.co/EGVAadTH1a'>,  
 favourite\_count: 1,  
 retweet\_count: 0,  
 created\_at: 2021-07-06T20:04:03.000Z }

* Query for phrase "Covid-19 vaccine" in Data Lake returned the same four records as relational database.
  + db.Collection0.find({"text":/Covid-19 vaccine/i}).sort({created\_at: -1}).pretty()
  + { \_id: ObjectId("60f9e95fffc0381c73f2c984"),  
     index: '20',  
     created\_at: 2021-07-20T10:20:00.000Z,  
     text: '51 percent of unvaccinated individuals think the COVID-19 vaccine contains a microchip <https://t.co/vz32hVzyWE'>,  
     retweet\_count: 0,  
     Location: 'US. Ohio' }

{ \_id: ObjectId("60f9e95fffc0381c73f2c98e"),  
 index: '30',  
 created\_at: 2021-07-20T10:20:00.000Z,  
 text: 'RT @DrIanWeissman: The U.S. will begin making its first COVID-19 vaccine shipments to Africa in the coming days, with the ultimate goal of \_',  
 retweet\_count: 0,  
 Location: 'NULL' }

{ \_id: ObjectId("60f9e95fffc0381c73f2c98a"),  
 index: '26',  
 created\_at: 2021-07-20T10:20:00.000Z,  
 text: 'RT @Breaking911: CCP NATIONAL HEALTH COMMISSION: COVID-19 vaccines must remain voluntary, local health authorities MAY NOT mandate the COVI \_',  
 retweet\_count: 0,  
 Location: 'NULL' }

{ \_id: ObjectId("60f9e95fffc0381c73f2c977"),  
 index: '7',  
 created\_at: 2021-07-20T10:20:00.000Z,  
 text: 'RT @GregSankey: COVID-19 vaccines are widely available, proven to be highly effective and, when people are fully vaccinated, we all have th \_',  
 retweet\_count: 0,  
 Location: 'McKinney, TX' }

* Data descriptions
  + \_id
    - Reference for the document
  + Index
    - Key within the original dataset but not over the whole Data Lake
  + Text
    - Text within the tweet
  + Retweet\_count
    - Count of retweets from other users
  + Favorite\_count
    - Count of favorites from other users
  + Created\_at
    - Timestamp of the tweet
  + Location
    - Location of the user when they sent the tweet