uOttawa

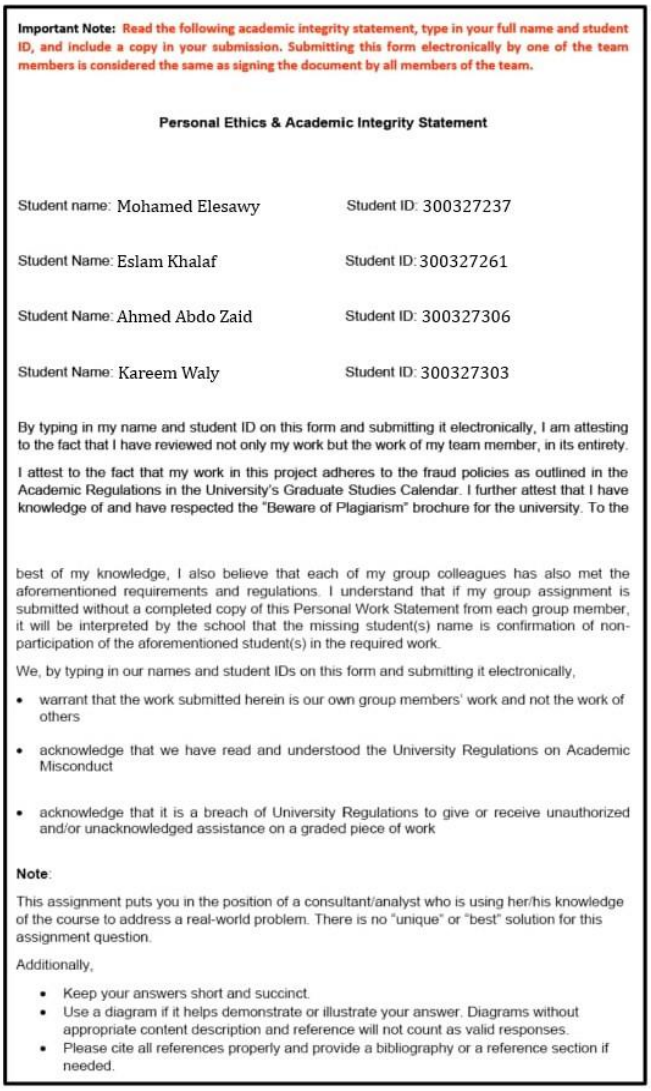
Faculty of Engineering

ELG 5166 – Cloud Analytics

Assignment #1

Group – 9

Mohamed Elesawy	300327237
Eslam Khalaf	300327261
Ahmed Abdo Zaid	300327306
Kareem Waly	300327303



Important Note: Read the following academic integrity statement, type in your full name and student ID, and include a copy in your submission. Submitting this form electronically by one of the team members is considered the same as signing the document by all members of the team.

Personal Ethics & Academic Integrity Statement

Student name: Mohamed Elesawy Student ID: 300327237

Student Name: Eslam Khalaf Student ID: 300327261

Student Name: Ahmed Abdo Zaid Student ID: 300327306

Student Name: Kareem Waly Student ID: 300327303

By typing in my name and student ID on this form and submitting it electronically, I am attesting to the fact that I have reviewed not only my work but the work of my team member, in its entirety.

I attest to the fact that my work in this project adheres to the fraud policies as outlined in the Academic Regulations in the University's Graduate Studies Calendar. I further attest that I have knowledge of and have respected the "Beware of Plagiarism" brochure for the university. To the best of my knowledge, I also believe that each of my group colleagues has also met the aforementioned requirements and regulations. I understand that if my group assignment is submitted without a completed copy of this Personal Work Statement from each group member, it will be interpreted by the school that the missing student(s) name is confirmation of non-participation of the aforementioned student(s) in the required work.

We, by typing in our names and student IDs on this form and submitting it electronically,

- warrant that the work submitted herein is our own group members' work and not the work of others
- acknowledge that we have read and understood the University Regulations on Academic Misconduct
- acknowledge that it is a breach of University Regulations to give or receive unauthorized and/or unacknowledged assistance on a graded piece of work

Note:

This assignment puts you in the position of a consultant/analyst who is using her/his knowledge of the course to address a real-world problem. There is no "unique" or "best" solution for this assignment question.

Additionally,

- Keep your answers short and succinct.
- Use a diagram if it helps demonstrate or illustrate your answer. Diagrams without appropriate content description and reference will not count as valid responses.
- Please cite all references properly and provide a bibliography or a reference section if needed.

Part 1: Questions

1. Describe briefly what a NoSQL database means. Select a NoSQL database (except MongoDB & Cassandra) and describe how this database can be used for the storage and management of big data.

A NoSQL database refers to 'not only SQL' that means it can store many types of data not only structured data but also semi structured data the availability is the most important factor for it. In relational databases, consistency and atomicity is the best important factor due to dealing with heavy inserting and updates queries then the performance of reading queries not good as needed because there are many joins that will slow down the performance, so in the big data needs massive reading queries that will not be suitable with relational databases, so NoSQL appears to solve massive reading big data problems.

Riak Database: is a particular kind of key-value store for a NoSQL database that can store any kind of data and does so over numerous nodes in what is referred to as a cluster. Data replication occurs in a cluster to support partition tolerance. This method of data replication ensures that this data is constantly accessible. Users of websites like social networking platforms or file-sharing programmers are an illustration of this. The data has been replicated among numerous nodes that extend across several servers, so it is safe in the case of a server crash, system failure, or other network anomaly. Once all nodes are successfully back online again, Riak starts to automatically allow reads and writes to and from the nodes that were not operational. In addition to extreme fault tolerance, Riak is highly scalable; it allows anyone to add new nodes with a few short commands and scale the size of the cluster to meet the data storage needs in a simple, linear, and predictable fashion, saving both time and money.

Or

AWS DynamoDB: is a fully managed key-value database offered by amazon web services, that delivers single-digit millisecond performance at any scale. you don't have to configure anything because it's a fully managed database, AWS manage scalability, configuration, replication, at rest encryption and hardware provisioning.

2. Investigate and describe one application of Big Data Analytics that was not described in

Class.

Big data analytics in organizational performance:
Large data samples can assist in identifying and leveraging business transformation and offer business-centric approaches and practices that give organizations or enterprises a competitive advantage by enabling them to enhance their current applications and daily operational decisions ad in creating new business values for the organization.

3. Briefly describe the transaction management features of Cassandra and MongoDB in the context of ACID vs. BASE properties.

Cassandra	MongoDB
• Cassandra is a NoSQL DB that can be categorized as a BASE database because of the fundamental availability, a soft state, and eventual consistency. • Cassandra is more efficient at the level of availability because it has multiple master nodes • Cassandra is a soft state because of the data's temporary inconsistency and change over time, it has a commit log, all data is written to the commit log as part of this backup technique to prevent data loss, and the system will change state without user intervention due to eventual consistency.	• MongoDB is a NoSQL DB supports ACID compliance at the document level.[3] • MongoDB supports multi-document ACID transactions on a single replica set.[3] • MongoDB supports availability with only one master nodes what make it not a best choice in case of availability.[3] • MongoDB isn't a soft state and eventual consistency because it's supported the consistency which mean that there will not be a lack of immediate consistency. [6]

- Cassandra does not immediately obligate consistency but updates data over time, and until it does, the data reads are possible even though they might not reflect reality.
- It possesses atomicity, isolation, and long-term durability.[1]
- It supports atomic(A), because inserting or modifying several columns in a row is considered a single write operation.[2]
- Cassandra cannot be described as an ACID-compliant database because the ACID lacks consistency(C) the concepts of referential integrity, foreign keys, and joins.[2]
- It supports isolation(I) Because it believes that writing to a row should be isolated to just the user who is writing and should not be visible to any other users until the current user has finished writing.[2]
- It supports durability(D) since all writes to a replica node are stored in memory and committed to disc before being declared successful.[2]

- The operation on a single document is atomic(A), since by using embedded documents to obtain relationships between data in a single document structure rather than normalizing across several documents and collections, the need for multi-document transactions is removed. [3]
- MongoDB supports consistency(C) the most recent value will be returned by any subsequent read after a write has finished.[3]
- MongoDB is isolation(I) since Data changes implemented during a transaction will not be visible outside of that transaction, until each transaction has finished processing, the data from other transactions cannot be accessed when several transactions are being carried out.[3]
- It supports durability(D) as it accommodates data from heterogeneous to homogeneous.[3]

4. You are working on a project that requires you to capture data from millions of IoT devices in people's homes. Each IoT device uploads a JSON document with the data elements required for analytics.

a) Identify potential NoSQL databases that you can capture data from the IoT Devices

o Document Database: such as MongoDB, it's a perfect solution for IOT app because it's power of storing data as JSON documents.[4]

o Time-Series Database: such as Timestream, it designed to hold time-oriented data, which makes it ideal for time-series-dependent machine learning models, IoT applications, and audit logs.[4]

o Columnar Database: such as Cassandra DB, considered as a good option because its power of ignoring the rows to directly query the columns. [4]

b) What are your design and analytics considerations and rationale behind your choice?

Since the IoT devices share huge streaming of data, a document database like MongoDB will be suitable because it does not require a schema, what make it more adaptable to changes, and it supports sharding to spread data across multiple servers to increase scalability and availability and can handle the flood of data from these devices, MongoDB scaled data either horizontally or vertically, and it natively supports time series data, which is typical of IoT architectures. In addition to the Column-oriented Database like Cassandra due to the fact that it may be used where analysis or grouping is being done on certain columns for the IoT applications, it's better than MongoDB in level of the availability because Cassandra Has multiple master nodes where MongoDB has only one master nodes, and the more master nodes there are in a cluster, the better the write speed (scalability), However, when it comes to consistency, MongoDB excels because you can query multiple nodes in a replica set and receive the same data, in contrast to Cassandra, which offers tunable consistency at the expense of performance.

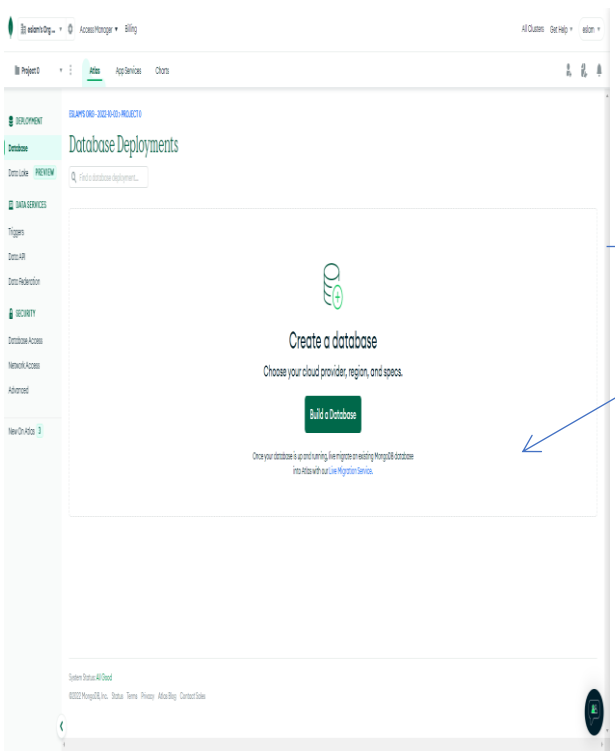
Finally, Time Series is an excellent choice too for collecting change events in the system, sensing, or storing device readings, each database of the above provides high scalability for IoT.[4] [5]

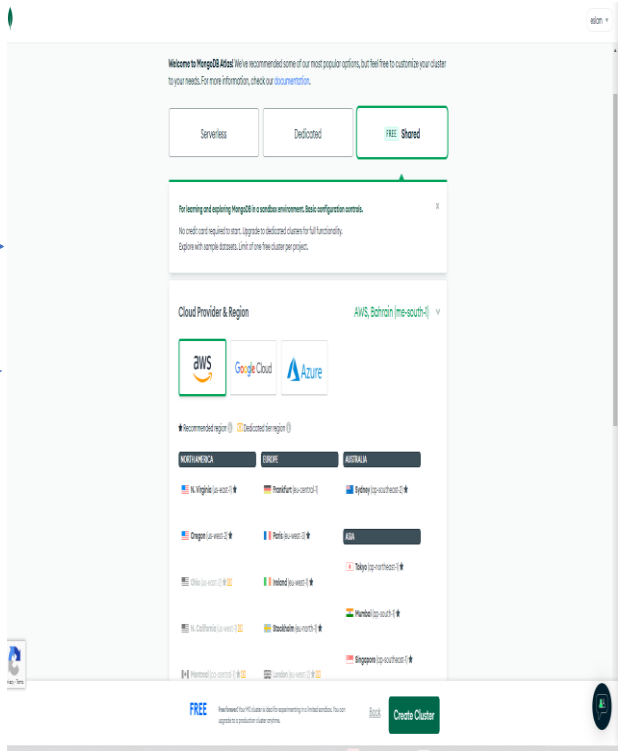
Part 2: NoSQL Labs

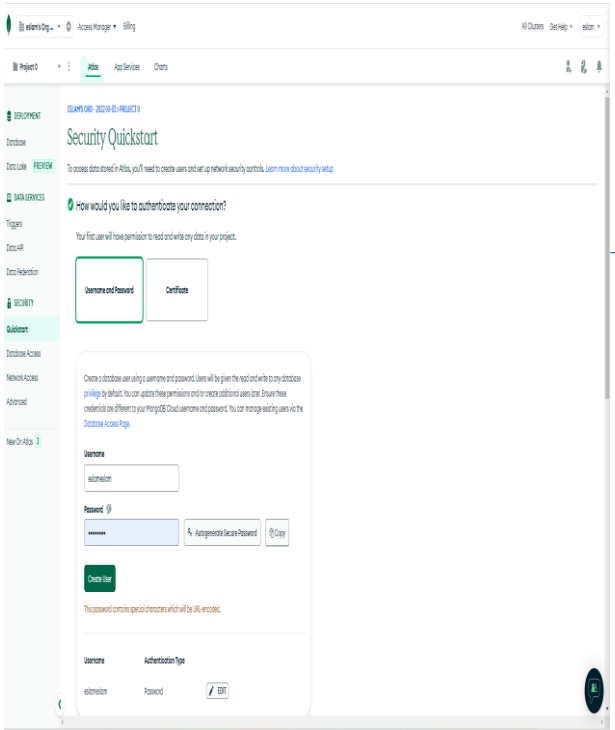
Setup (Please show evidence of your setup with screenshots)

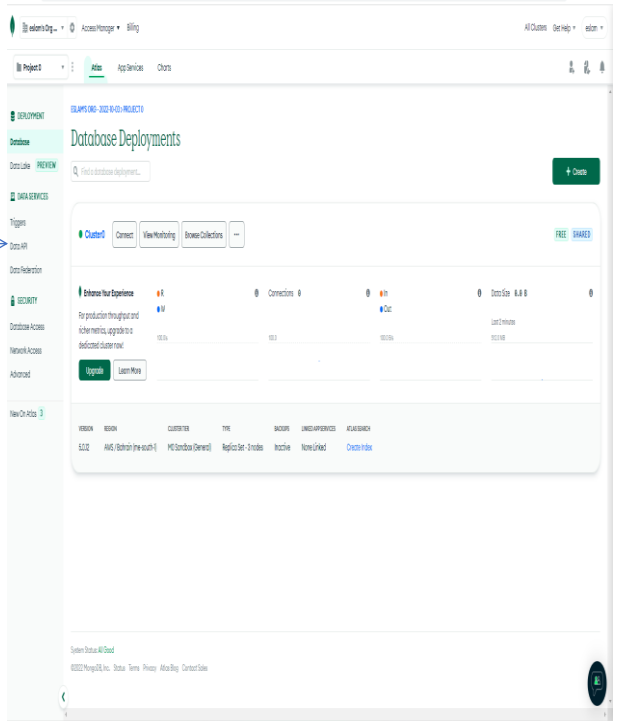
1) MongoDB Lab

1. Set an account on MongoDB Atlas - https://cloud.mongodb.com







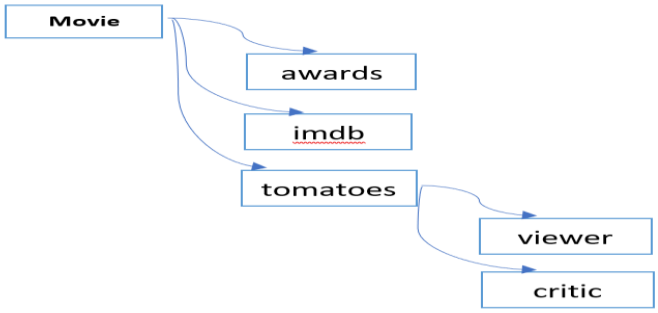


Answers:

1. Briefly describe the movies database document model.

After loading Sample Netflix Movies Database, we noticed documents contain the following attributes, but these attributes not found in all documents but these attributes we noticed them by showing some documents in our database:

This is the data model for our document that represents that the movie document has many (awards, imdb, tomatoes) and each tomatoes has many (viewer, critic)



- "_id" with type "ObjectId" that represent the document ID of each document
- "plot" with type "String"
- "genres" with type "Array"
- "runtime" with type "Integer"
- "cast" with type "Array"
- "num_mflix_comments" with type "Integer"
- "poster" with type "String"
- "title" with type "String"
- "fullplot" with type "String"
- "languages" with type "Array"
- "released" with type "Timestamp"
- "directors" with type "Array"
- "rated" with type "String"
- "awards" with type "Object" that object contain the following ("wins" with type "Integer", "nominations" with type

"Integer" and "text" with type "String")

- "lastupdated" with type "String"
- "year" with type "Integer"
- "imdb" with type "Object" that object contain the following ("rating" with type "Double", "votes" with type "Integer" and " id" with type "Integer")
- "countries" with type "Array"
- "type" with type "String"

•“tomatoes” with type “Object” that object contain the following (“viewer” with type “Object” that object contain the

following (“rating” with type “Double”, “numReviews” with type “Integer” and “meter” with type “Integer”)

“critic” with type “Object” that object contain the following (“rating” with type “Double”, “numReviews” with type “Integer” and “meter” with type “Integer”)

“rotten” with type “Integer” and

“fresh” with type “Integer” and

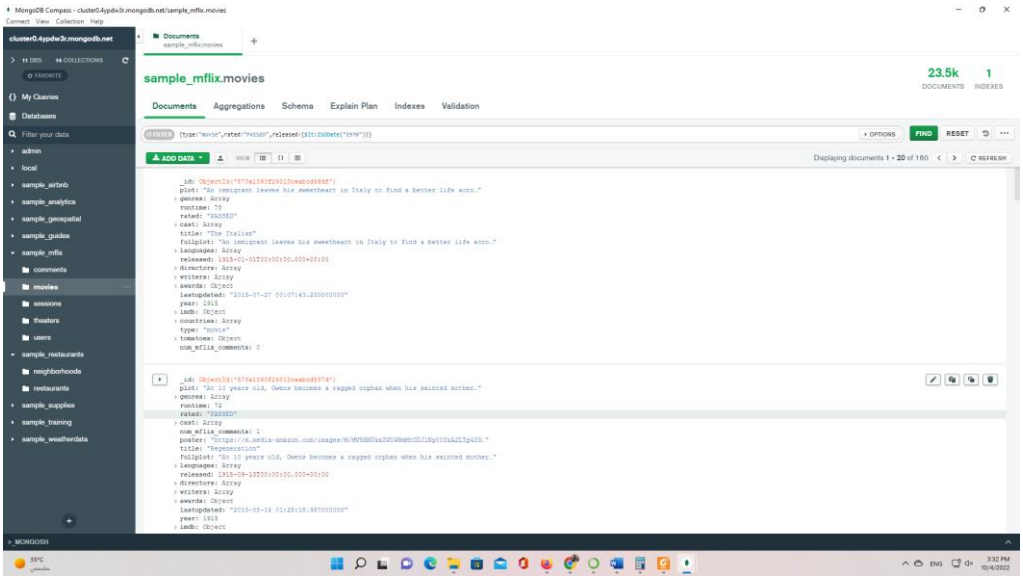
“dvd” with type “Timestamp” and

“lastUpdated” with type “Timestamp”).

All of this represents each document in Netflix Movies Database.

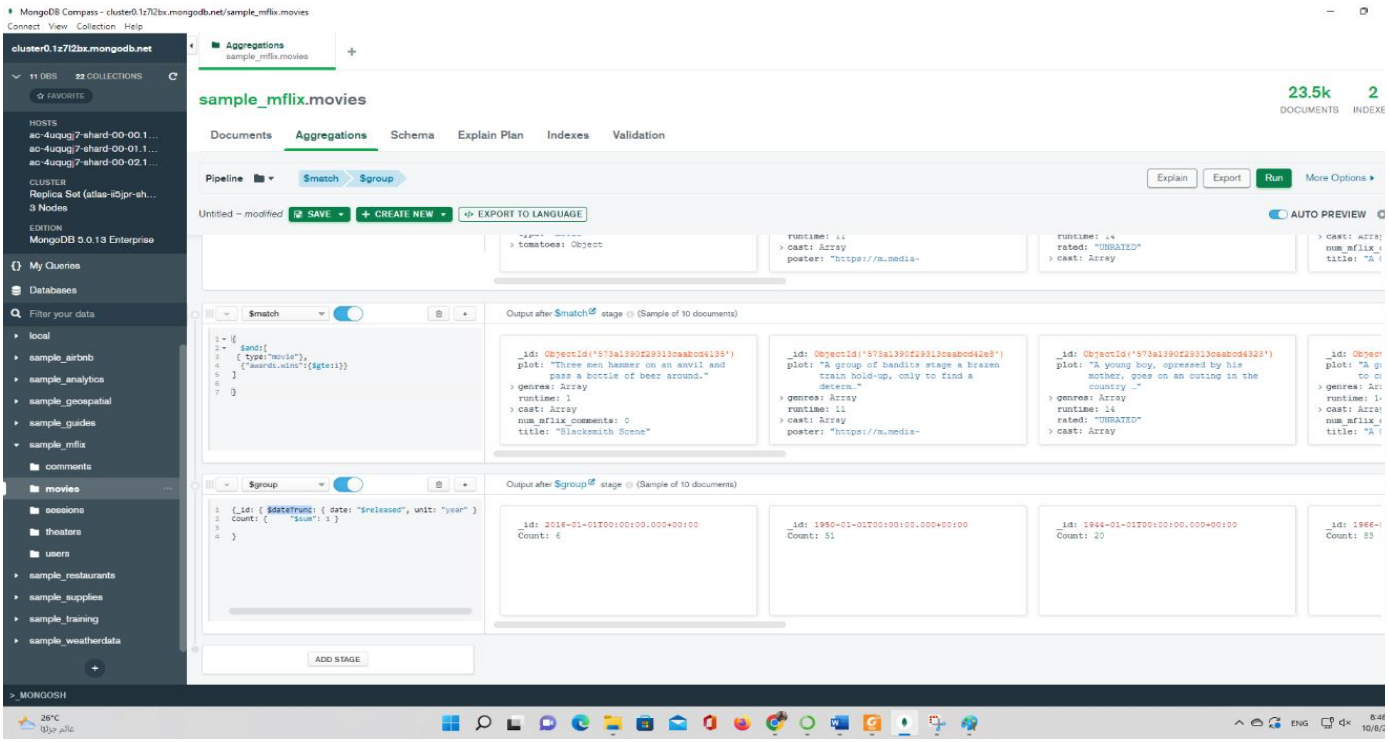
2. **Filter the documents for type “movies” that are releasedbefore 1970 and rated as “PASSED”.**

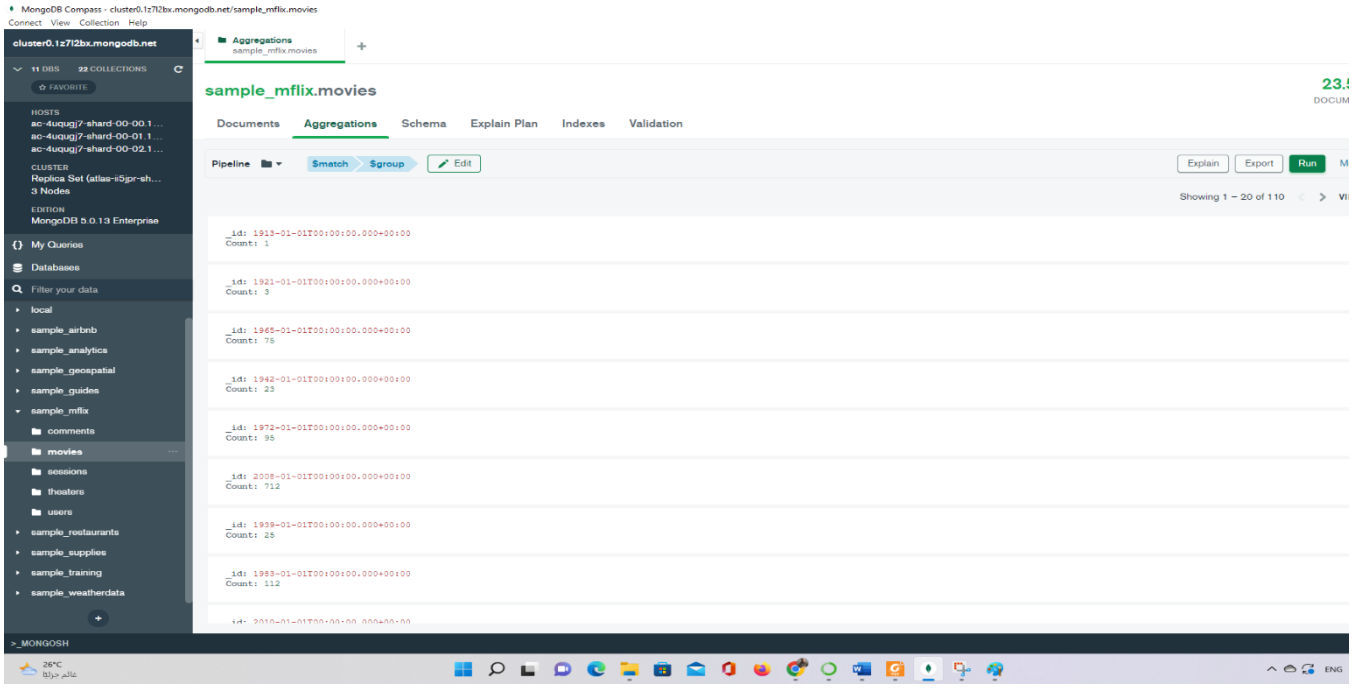
{type:"movie",rated:"PASSED",released:{$lt:ISODate("1970")}}



2. **Build an Aggregation Pipeline that shows all movies (“type”) that have won at least one award and provide the results in release year aggregate count.**

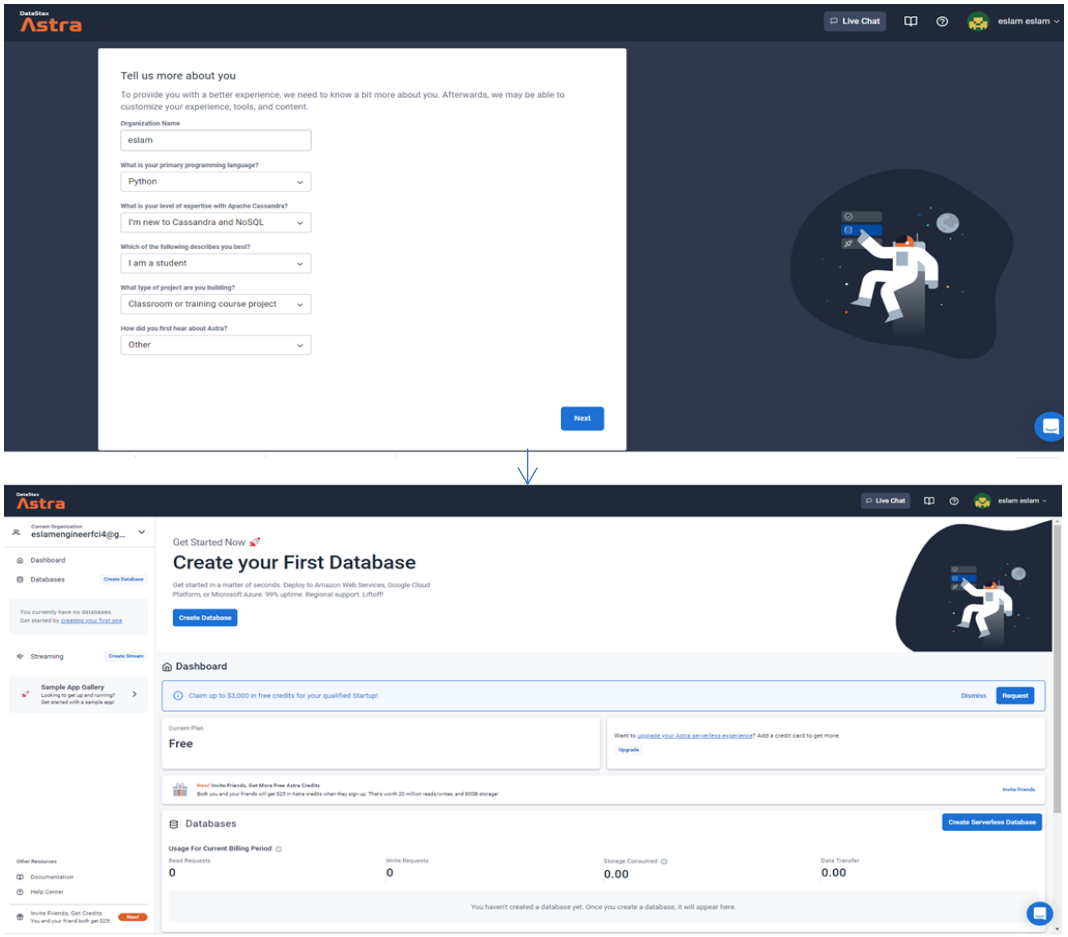
```
{
  $and:[
 { type:"movie"},
  {"awards.wins":{$gte:1}}
]

}

{_id: { $dateTrunc: { date: "$released", unit: "year" } },
Count: {     "$sum": 1 }

}
```

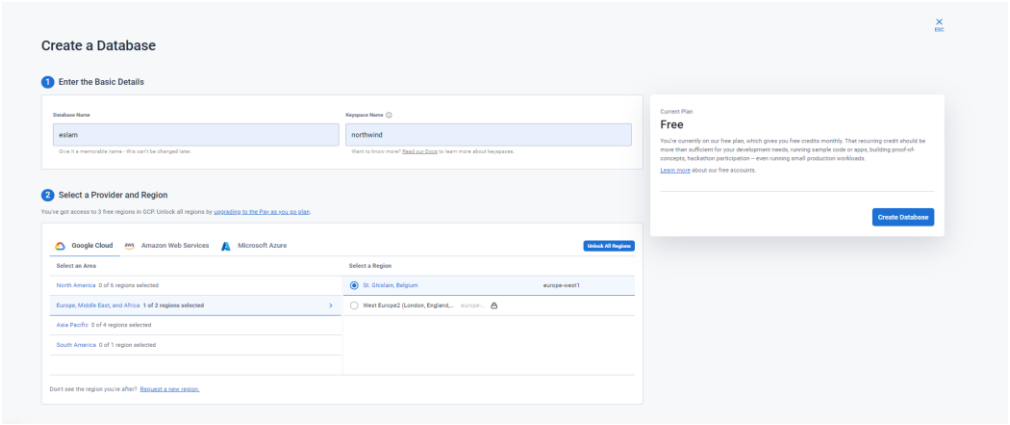

MongoDB Compass - cluster0.1z7l2bx.mongodb.net/sample_mflix.movies
cluster0.1z7l2bx.mongodb.net
sample_mflix.movies
23.5k DOCUMENTS
2 INDEXES
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
Pipeline
$match
$group
Explain
Export
Run
More Options
SAVE
CREATE NEW
EXPORT TO LANGUAGE
AUTO PREVIEW
ADD STAGE
>_MONGOSH

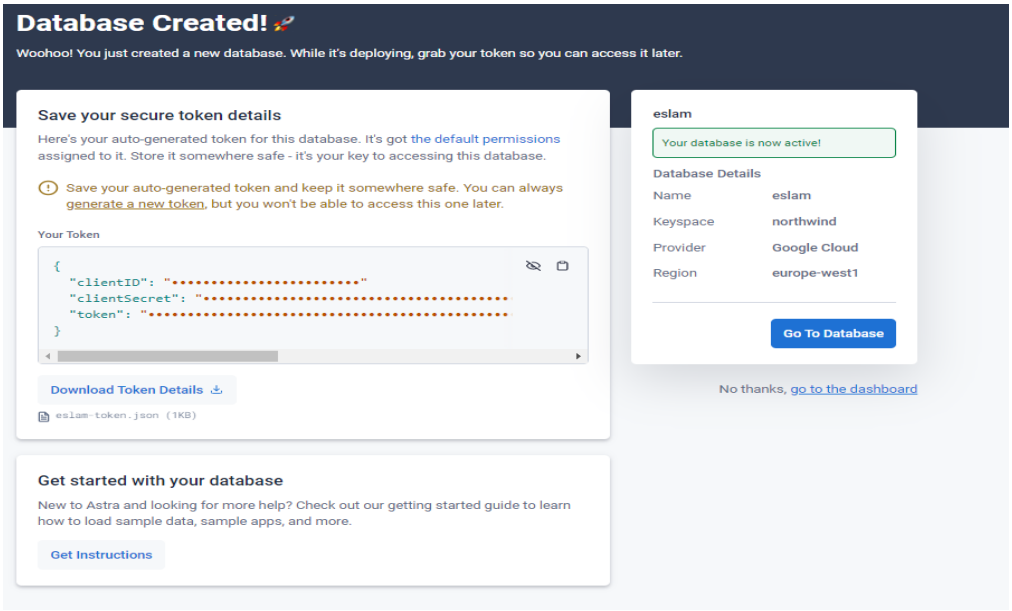

MongoDB Compass - cluster0.1z7l2bx.mongodb.net/sample_mflix.movies
cluster0.1z7l2bx.mongodb.net
sample_mflix.movies
Documents
Aggregations
Schema
Explain Plan
Indexes
Validation
Pipeline
$match
$group
Edit
Explain
Export
Run
Showing 1 – 20 of 110
>_MONGOSH

2) Cassandra Lab

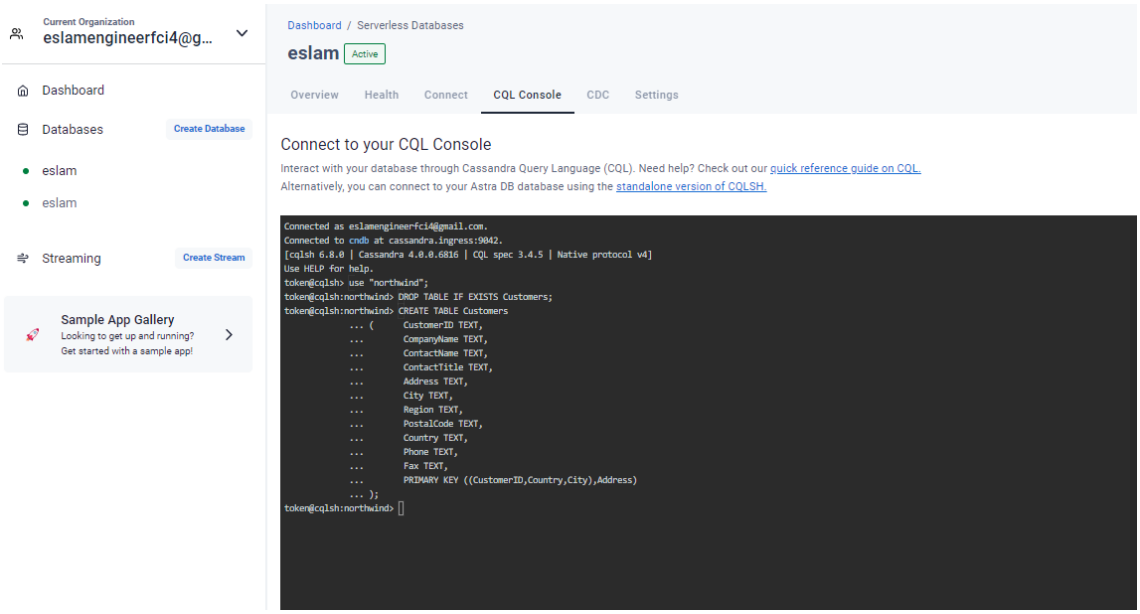
Using DataStax Astra Cassandra-as-a-Service (https://astra.datastax.com/)



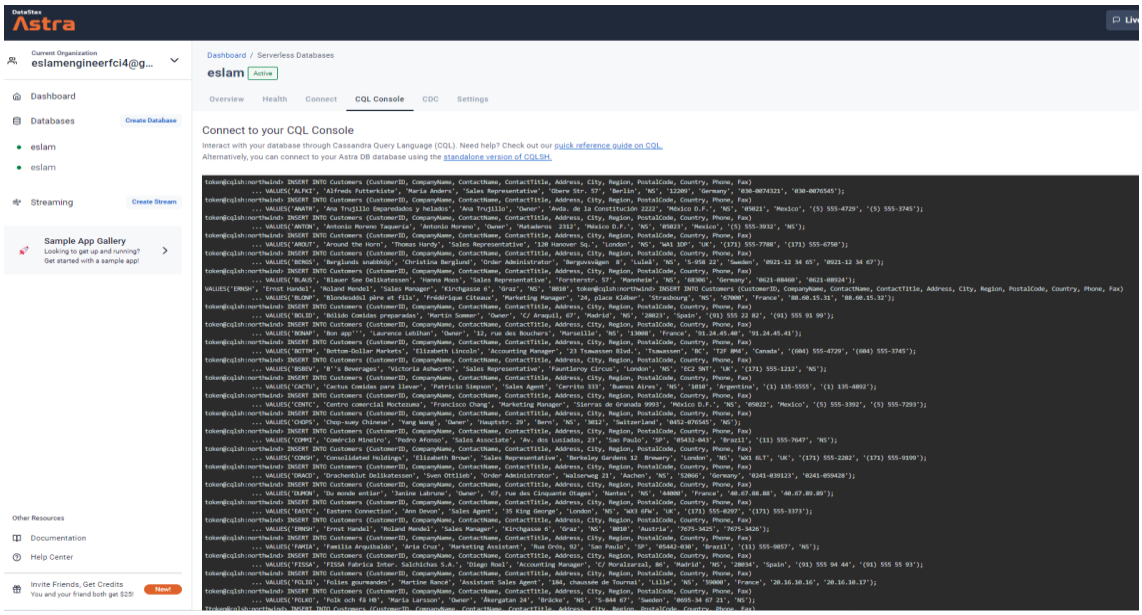




Create the customer tables (attached SQLite definition will serve as a guide) Review the questions in the queries section below and create one or more tables that partition and clusters data so these queries will execute without using Cassandra “ALLOW FILTERING” that scans all partitions.



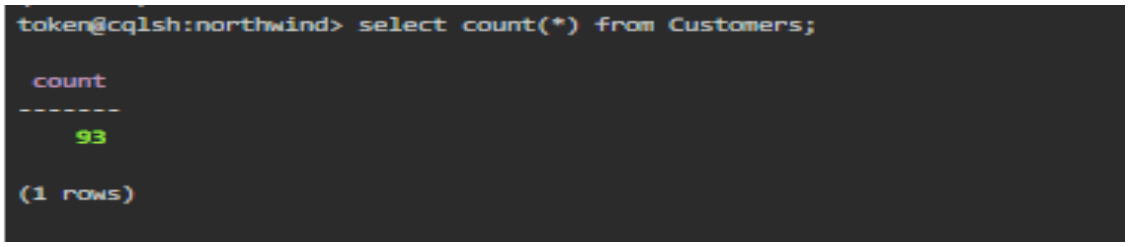
Load the attached data into your table(s) using the insert statements (minor modifications may be needed if your definitions include multiple tables). Please include screens counts after loading your data.



Select * from Customers limit 10; (show 10 records from customers table)

```
token@cqlsh:northwind> select * from Customers limit 10;

 customerid | country   | city           | address                    | companyname                    | contactname      | contacttitle        | fax            | phone          | postalcode | region
------------+-----------+----------------+----------------------------+--------------------------------+------------------+---------------------+----------------+----------------+------------+-------------
      LINOD | Venezuela | I. de Margarita |      Ave. 5 de Mayo Porlamar |              LINO-Delicateses | Felipe Izquierdo |               Owner |    (8) 34-93-93 |    (8) 34-56-12 |       4980 | Nueva Esparta
      DUMON |    France |         Nantes | 67, rue des Cinquante Otages |            Du monde entier |  Janine Labrune |               Owner |    40.67.89.89 |    40.67.88.88 |      44000 |          NS
      HUNGO |   Ireland |           Cork |          8 Johnstown Road |       Hungry Owl All-Night Grocers | Patricia McKenna |   Sales Associate |       2967 3333 |       2967 542 |        NS |      Co. Cork
      BOLID |     Spain |         Madrid |           C/ Araquil, 67 |          Bólido Comidas preparadas |   Martín Sommer |               Owner | (91) 555 91 99 | (91) 555 22 82 |      28023 |          NS
      PRINI |  Portugal |         Lisboa |      Estrada da saúde n. 58 |          Princesa Isabel Vinhos | Isabel de Castro | Sales Representative |          NS |    (1) 356-5634 |       1756 |          NS
      ANATR |    Mexico |    México D.F. | Avda. de la Constitución 2222 | Ana Trujillo Emparedados y helados |  Ana Trujillo |               Owner |    (5) 555-3745 |    (5) 555-4729 |      05021 |          NS
      LONEP |       USA |       Portland |          89 Chiaroscuro Rd. |          Lonesome Pine Restaurant |  Fran Wilson |       Sales Manager | (503) 555-9646 | (503) 555-9573 |      97219 |          OR
      VINET |    France |          Reims |          59 rue de l'Abbaye |          Vins et alcools Chevalier |  Paul Henriot |  Accounting Manager |    26.47.15.11 |    26.47.15.10 |      51100 |          NS
      LETSS |       USA |  San Francisco |       87 Polk St. Suite 5 |                 Let's Stop N Shop |  Jaime Yorres |               Owner |          NS | (415) 555-5938 |      94117 |          CA
      QUEDE |    Brazil | Rio de Janeiro |     Rua da Panificadora, 12 |                  Que Delícia | Bernardo Batista |  Accounting Manager | (21) 555-4545 | (21) 555-4252 |   02389-673 |          RJ

(10 rows)
token@cqlsh:northwind>
```

Select count (*) from Customers;



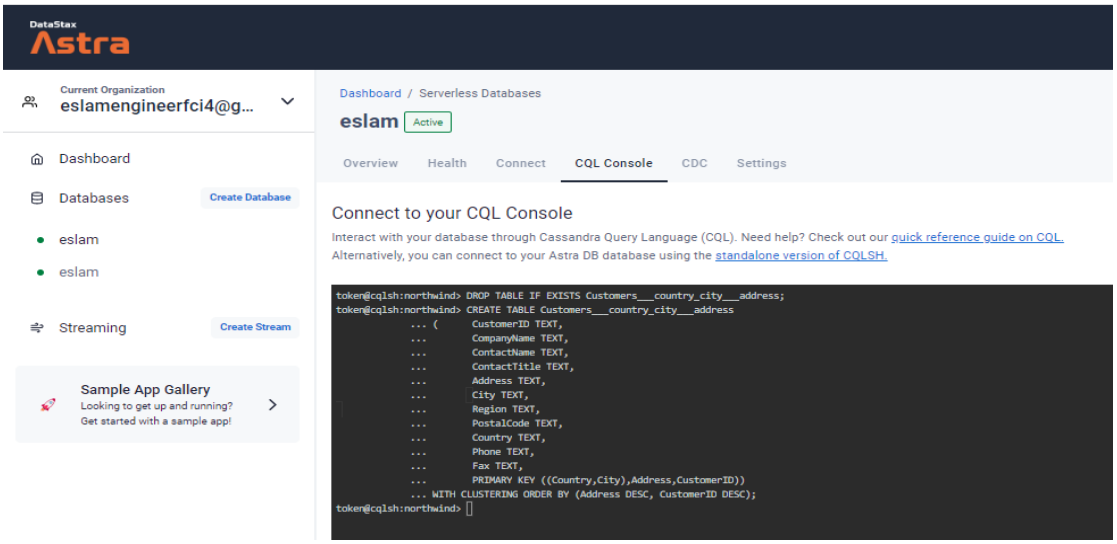
```
token@cqlsh:northwind> select count(*) from Customers;

 count
-------
    93

(1 rows)
```

Queries:

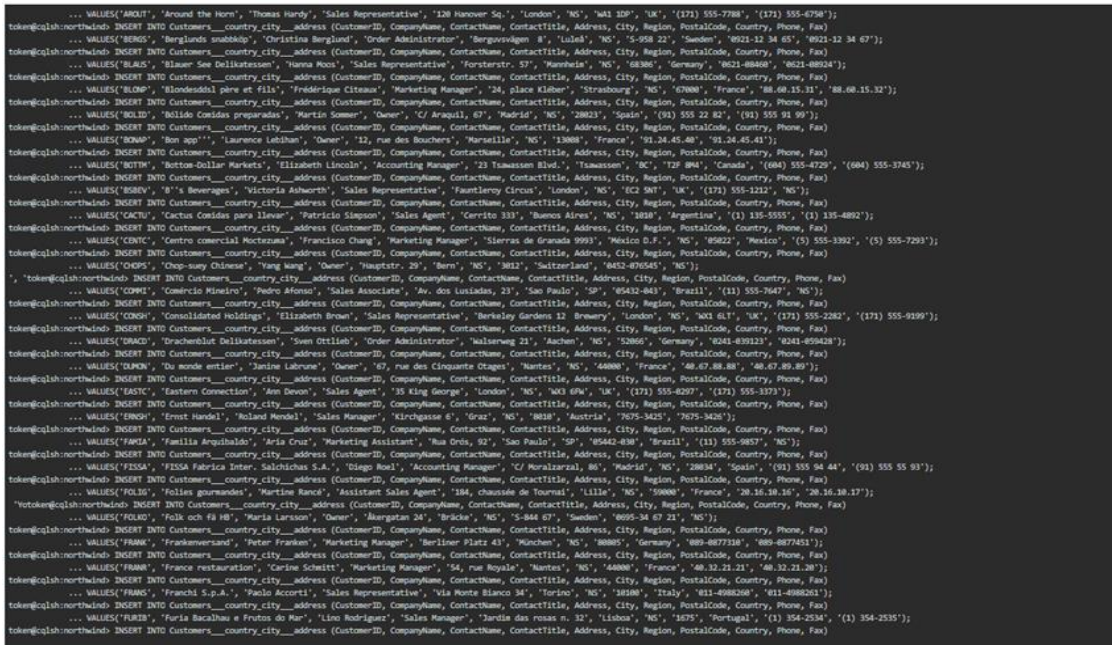
1) Provide the query and the results (screenshots and a copy of your query) that show the customers from Rio de Janeiro, Brazil ordered by their addresses.

Our first table is Customers___country_city___address that represent our first query

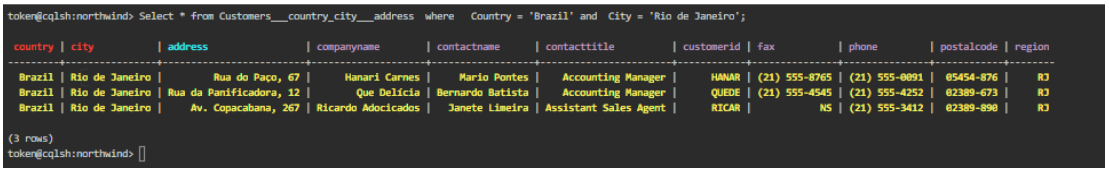
In our first table we used (country, city) as partition keys and (address, customerId) as cluster keys because these four values will make a unique value for each record



There aren't any data yet so we will insert our data to our table that will be used in first query



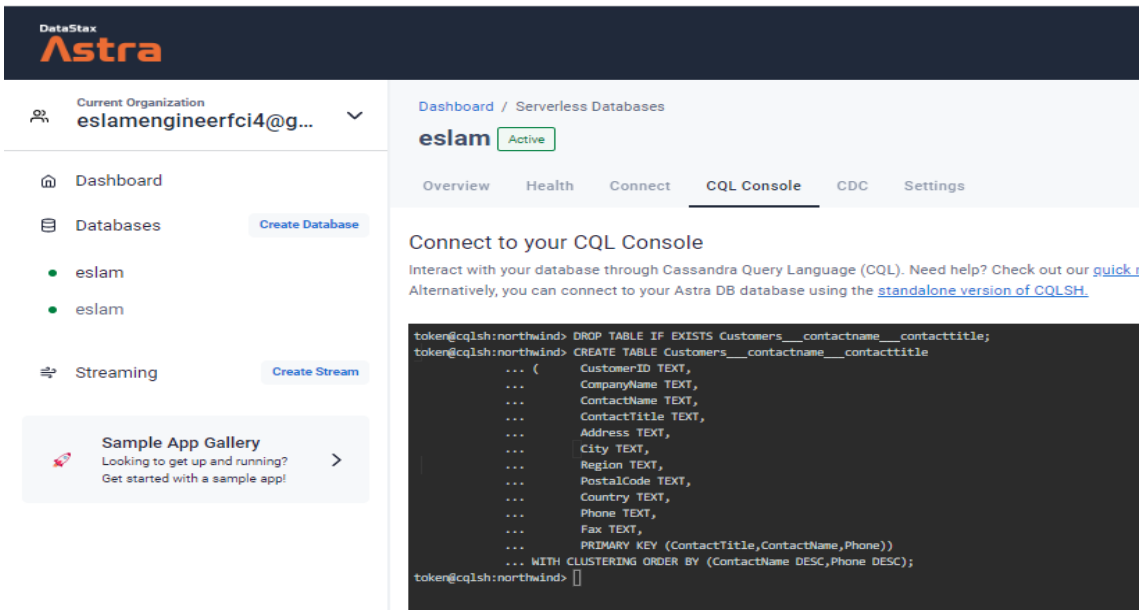
Select * from Customers___country_city___address where Country = 'Brazil' and City = 'Rio de Janeiro';

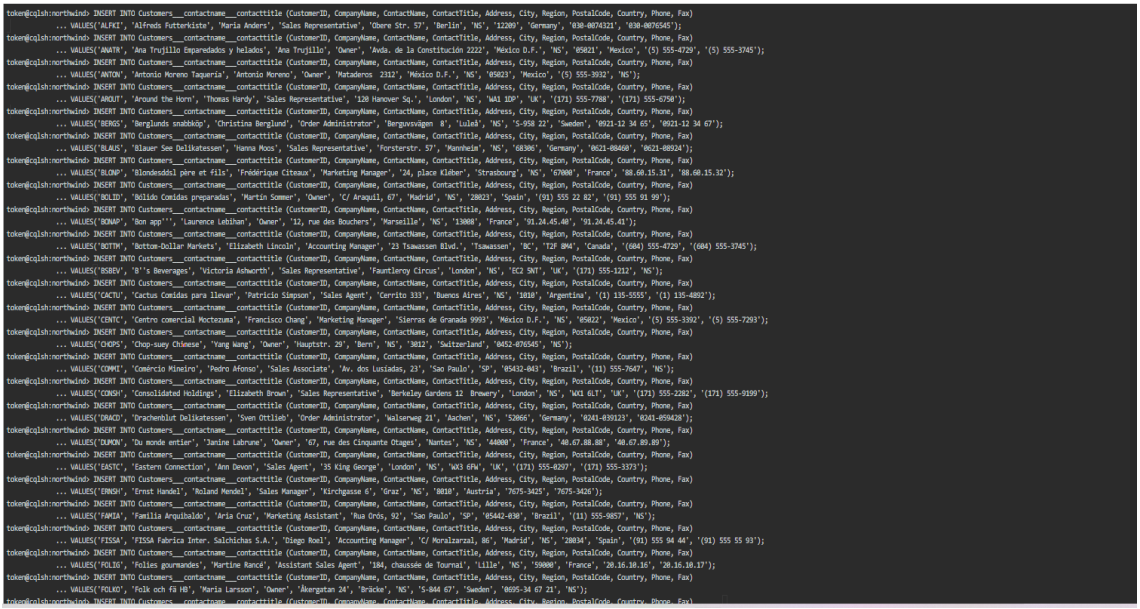


2) Provide a list of customers that are in the Sales Manager role without forcing the scan of all partitions across all databases. The result should be ordered by their names.

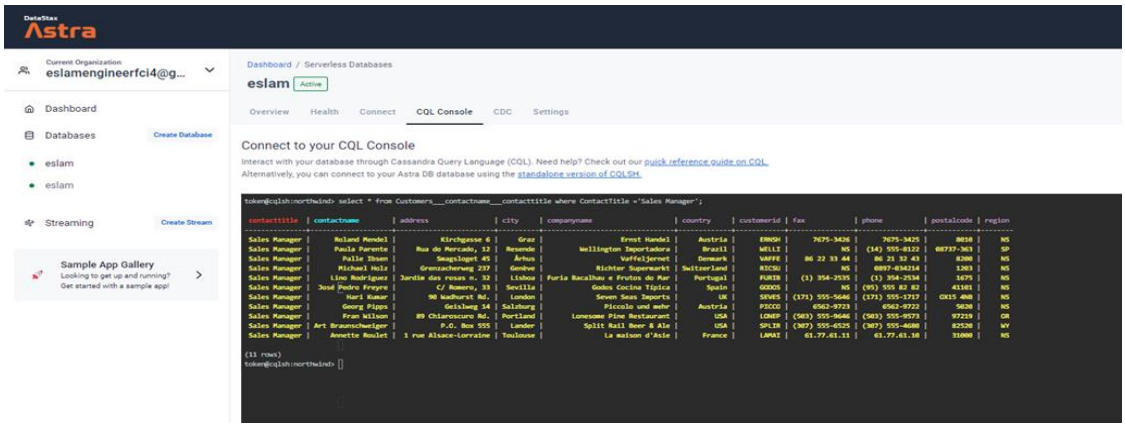
Our second table is Customers___contactname___contacttitle that represent our second query and insert data to it

In our second table we used (ContactTitle) as partition key and (ContactName, Phone) as cluster keys because these three values will make a unique value for each record





select * from Customers___contactname___contacttitle where ContactTitle ='Sales Manager';



References

[1] https://www.yugabyte.com/blog/apache-cassandra-lightweight-transactions-secondary-indexes-tunable-consistency/

[2]https://docs.datastax.com/en/cassandraoss/2.2/cassandra/dml/dmlTransactionsDiffer.html

[3] https://studio3t.com/whats-new/mongodb-acid-properties/

[4] https://www.verypossible.com/insights/nosql-for-iot-development-how-to-choose-the-right-database

[5] https://www.openlogic.com/blog/cassandra-vs-mongodb

[6] https://www.geeksforgeeks.org/acid-model-vs-base-model-for-database/